MULTIPLICATIVELY PERTURBED LEAST SQUARES FOR DIMENSION REDUCTION

by
M.P. Martin

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May, 2021

Abstract

Dimension reduction is a crucial aspect of modern data science, offering computational efficiency, insight into the structure of problems, and increased accuracy for downstream regression problems. According to a well-known result in approximation theory, the mean squared error of a non-parametric regression problem is not guaranteed to decrease faster than $N^{-\frac{2p}{2p+D}}$, where N is the number of samples, p a smoothness parameter of the problem, and D the dimension of the inputs. This slow rate is due to the so-called "Curse of Dimensionality," in which samples in high-dimensional domains are exponentially likely to be well isolated from each other. These concerns motivate research into algorithms to determine intrinsic structure to the functions being regressed, as any reduction in D yields an exponential improvement in the lower bound of sample complexity. Even in parametric settings, large D increases computational complexity and hinders the ability to find useful parameter values.

In this thesis, we discuss various existing methods of dimension reduction and introduce our own: Multiplicatively Perturbed Least Squares (MPLS). We provide a theoretical analysis of MPLS that proves it achieves the optimal convergence rate of $N^{-1/2}$ for a broad class of functions, up to logarithmic factors. This theoretical analysis is supplemented by a series of experimental results, in which MPLS performs better or comparable to existing dimension reduction algorithms.

Thesis Readers

Fei Lu

Assistant Professor

Department of Mathematics at

Johns Hopkins Krieger School of Arts and Sciences

Mauro Maggioni

Bloomberg Distinguished Professor

Department of Mathematics & Department of Applied Mathematics and Statistics at

Johns Hopkins Krieger School of Arts and Sciences and Whiting School of Engineering

Acknowledgements

My first and dearest thanks go to my parents, Mike Martin and Pat McAlarnen, and my aunt, Eileen McAlarnen, who have supported me at every stage of my education. They have modeled a standard of intelligence, curiosity, adventure, and respect as adults that have been invaluable as the primary figures in my life.

I would also like to thank my advisor, Mauro Maggioni, for his guidance, imagination, and patience over the course of this program. Countless roadblocks in conducting this research have been cleared through simple comments of his in our meetings.

In a quite literal sense, this would not have been possible without the Math Department administrative staff, especially Sabrina Raymond. Through distributing information about due dates, graduate student opportunities, and when the leftover wine and cheese food was put out, they helped make sure I knew what was going on.

My time at Johns Hopkins was made infinitely more enjoyable thanks to the friends I have made along the way. Foremost among these are Daniel Fuentes-Keuthan and tslil clingman, whom I have shared innumerable cherished memories with. Ben Dees, Ben Peak, David Myers, and Jeff Marino fostered a fun atmosphere inside and outside the math department that made it enjoyable to show up to campus to work (in normal times)—thank you. Dmitri Bobrovnikov, Jen Torres, and Mark Petersen were the mainstays in a trivia team that was my primary social group during my first year here and really helped me get my footing in being a graduate student. Many others have had a hand in keeping me sane here over the years, including Chris Kauffman, Dan

Ginsberg, and Jane Lutken.

I am also blessed to have a strong network of friends outside of Johns Hopkins. My friends from my undergraduate at the University of Arizona, including Alexa Bautista, Jaggar Henzerling, Jeff O'Hara, Riley Tabar, and Zach Van Uum have been irreplaceable and have kept me well supplied with terrible movies and obnoxious video games. Jacqui Oesterblad, with her sharp mind and admirable moral compass, has been a constant reference in my mind as my personal philosophies have evolved. I am forever grateful to Bryan Fike for his presence and assistance during a particularly tough time in my life; the mental resiliency and self-confidence I discovered through that period have made completing this degree possible. Through American Model United Nations I have met so many people—including Alex Middlewood, Jackie Whitt, Nate Bode, Nate Ritsema, Paul Kruchoski, and Zach Chastain—from different walks of life who have given me knowledge and experiences I would never have gotten elsewhere.

Finally, I would like to thank the institutions that supported my path to this point. Las Brisas Elementary School, Hillcrest Middle School, Sandra Day O'Connor High School, and the University of Arizona all offered unique experiences—educational, extra-curricular, and social—without which I am sure I would never have reached this point.

Thank you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Machine learning involves tasks in which one has collections of data pairs $(\mathbf{x}_i, y_i)$, themselves realizations[1] of some random variables $\mathbf{X}$ and Y, with the hope of there being some structure to the data so that novel $(\mathbf{x}, y)$ pairs can be estimated well using only the $\mathbf{x}$ value. The $\mathbf{x}$ values will typically be described as living in some subset $\Omega \subset \mathbb{R}^D$, with y living in $\mathbb{R}$; this occasionally involves re-labeling the data[2]. There are three broad classes of machine learning problems that these tasks tend to fall in:

- Unsupervised learning, in which the problem is identifying structure among the $\mathbf{x}_i$, with the y_i representing cluster labels, outlier status, or other qualities. These "true" y values are typically unobserved, even in the training data, with performance instead measured based on expected properties of the y's, such as minimizing cluster diameter or maximizing log-likelihood.

- Classification, in which typically the y values are contained in a discrete label set $\mathcal{L}$ and the problem is replicating the *Bayes' Classifier*, $f(\mathbf{x}) = \arg\max_{k \in \mathcal{L}} \mathbb{P}[Y = k | \mathbf{X} = \mathbf{x}]$, often optimizing horseshoes-and-hand-grenades

[1] Not necessarily independent or identically distributed, but we will make that assumption soon.

[2] We might identify non-numeric sets with discrete sets, for example in text processing where $\mathbf{x}$-values are often vectors indicating whether a word appears in the document, or in image classification where the y-values of $\{\text{Cat}, \text{Not Cat}\}$ are treated as $\{0, 1\} \subset \mathbb{R}$

measurements of accuracy and precision (such as the F-score) that weight all errors equally[3].

- Regression, in which the y values are typically not discrete but instead related to $\mathbf{x}$ via a function $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = f(\mathbf{x})$, often written $Y = f(\mathbf{X}) + \varepsilon$ where ε is some mean-zero random variable, and the problem is to minimize the total difference between the observed y and the predicted $\hat{y}$, often looking at the L^p norm of the difference, in particular the Mean Squared Error (MSE) or squared L^2 norm.

The lines between these three cases are often blurred—one popular algorithm for classification is logistic regression, which eponymously is a regression algorithm—however we will focus on the problem of regression, which attempts to answer questions like

- Given data about the orbital characteristics of a planet (mass, semimajor axis, etc), predict its period[4]
- Given polling data and past election results, predict the outcome of an election
- Given satellite images of cropland, predict crop yield ([GKKW02], p.5)
- Given neighborhood quality information, predict house price ([HR78])

Sometimes the results of regression are interesting in their own right—the ability to forecast an election helps calibrate expectations beyond "punditry" ([SMP96]). However, this is not the only potential value: in [HR78], the prediction of house values is then used to infer the impact of air quality on housing prices; in the crop yield

[3] In other words, there is no partial credit for being close, as in the saying "'almost' only counts in horseshoes and hand grenades"

[4] This is Kepler's Third Law, whose 1.5 power relationship Kepler famously "guessed" from empirical data.

problem, further analysis was done to determine what bands (colors) of light in the images were most relevant for predicting yields.

In this thesis, we will follow certain notational conventions. **Boldface** variables will represent vectors, with components $\mathbf{x} = \left(x^1, \ldots, x^D\right)$; β, η, and ν will also be vectors. Capital letters will denote matrices, random variables, and important numerical constants. Calligraphic letters, for example $\mathcal{S}$, will denote sets. The notation $\|\cdot\|$ without subscript represents the Euclidean norm if the argument is a vector, the L^2 norm if the argument is a function, or the spectral norm if the argument is a matrix. A hat accent, for example $\hat{f}$, indicates a finite-sample estimate intended to approximate an unknown object f. The function log without a subscript denotes the logarithm with base e. The symbols $\sim$ and $\perp\!\!\!\perp$ in the context of random variables mean "distributed like" and "is independent of", respectively.

1.2 Problem Setup for Regression

Our problem setup is as follows:

- $\mathbf{X}$ lives in a probability space $(\mathbb{R}^D, \phi)$, Y is a random variable not independent of $\mathbf{X}$ such that $\mathbb{E}\left[Y^2\right] < \infty$
- We have N pairs of samples $(\mathbf{x}_i, y_i) \sim \mathbf{X} \times Y$ drawn with respect to the product probability distribution
- We wish to find a measurable function $\hat{f}$ that minimizes the mean squared error (MSE)[5],

$$\left\|Y - \hat{f}(\mathbf{X})\right\|_{MSE}^2 := \mathbb{E}_{\mathbf{X},Y}\left[(Y - \hat{f}(\mathbf{X}))^2\right]$$

The process of finding such an $\hat{f}$ is called a regression algorithm:

[5] We shortly will be looking at expectations of this MSE, hence why $\|\cdot\|^2$ notation is used here in place of $\mathbb{E}\left[\cdot^2\right]$

Definition 1 (Regression)**.** *A regression algorithm ψ_N is a function that takes a training dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and returns a function $\psi_N(\mathcal{S}) = \hat{f}_N : \mathbb{R}^D \to \mathbb{R}$.*

Of course, the regression algorithm can only hope[6] to minimize the empirical MSE:

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2.$$

This is substantively a different question than minimizing the true MSE. Many regression algorithms, such as ridge regression ([HK70]), will minimize a quantity related to the empirical MSE but that incorporates "regularization" parameters that help ensure better performance on out-of-sample data.

The function $f(\mathbf{x}) = \mathbb{E}[Y\,|\mathbf{X} = \mathbf{x}]$ minimizes the MSE; we will call this function the *regression function*. A benefit of choosing to minimize the MSE, instead of another potential loss function, is that we can write the MSE of an estimate $\hat{f}$ as a sum of two terms:

$$\begin{aligned}
\left\|\hat{f} - Y\right\|_{MSE}^2 &= \mathbb{E}\left[(\hat{f}(\mathbf{X}) - Y)^2\right] \\
&= \mathbb{E}\left[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2\right] + \mathbb{E}\left[(f(\mathbf{X}) - Y)^2\right] \\
&\quad - 2\mathbb{E}\left[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))(f(\mathbf{X}) - Y)\right] \\
&= \mathbb{E}\left[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2\right] + \mathbb{E}\left[(f(\mathbf{X}) - Y)^2\right] \\
&\quad - 2\mathbb{E}\left[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))\mathbb{E}\left[f(\mathbf{X}) - Y\,|\mathbf{X}\right]\right] \\
&= \mathbb{E}\left[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2\right] + \mathbb{E}\left[(f(\mathbf{X}) - Y)^2\right] \qquad (1.1)
\end{aligned}$$

Since both summands are positive, we see that f does truly minimize the MSE. Additionally, this decomposition shows that the MSE of an estimate $\hat{f}$ depends only on the error between $\hat{f}$ and f. The MSE is bounded away from zero by a term representing the variance of the noise, $\mathbb{E}[(f(\mathbf{X}) - Y)^2]$.

[6] A regression algorithm is not *required* to minimize the empirical MSE, or to minimize anything. A *good* regression algorithm will likely attempt to do something along these lines, but in this theoretical analysis we won't assume any aspect of the behavior of the algorithms.

Given various regression algorithms, we would like to talk about how well they do at reproducing the regression function. There, again, are a number of ways of measuring the difference between two functions, however we see in (1.1) that minimizing the L^2 distance from f, $\mathbb{E}\left[\left(\hat{f}-f\right)^2\right]$, also satisfies our original goal of minimizing the MSE, and so we will use this metric.

Our goal of finding f is hampered by the fact that viewing only finitely many datapoints will never be enough to determine f exactly: consider the δ-memorization function $m_{\mathcal{S}}^{\delta}$

$$m_{\mathcal{S}}^{\delta}(\mathbf{x}) = \begin{cases} 0 & \|\mathbf{x}-\mathcal{S}\| \leq \frac{\delta}{2} \\ 1 & \|\mathbf{x}-\mathcal{S}\| > \delta \end{cases}$$

The behavior of $m_{\mathcal{S}}^{\delta}$ is purposefully left undefined between $\delta/2$ and δ and will be discussed later. Importantly, with $\delta = 0$, no countable amount of data will allow us to distinguish f from $f + m_{\mathcal{S}}^{0}$, however we would also have to be *extremely* unlucky[7] to see exactly the data corresponding to $\mathcal{S}$. Because of this, we ask for asymptotic convergence in expectation, which we will call *consistency*:

Definition 2. *A regression algorithm ψ_N is consistent if*

$$\lim_{N\to\infty} \mathbb{E}_N\left[\|f(\mathbf{X}) - \psi_N(\mathcal{S})(\mathbf{X})\|^2\right] = 0$$

where the expectation is taken with respect to samples $\mathcal{S}$ of N i.i.d. datapoints $(\mathbf{x}_i, y_i)$. We say ψ_N is N^{α}-consistent if it is consistent and, for any[8] $\varepsilon > 0$,

$$\lim_{N\to\infty} N^{-2\alpha+\varepsilon}\mathbb{E}_N\left[\|f(\mathbf{X}) - \psi_N(\mathcal{S})(\mathbf{X})\|^2\right] < \infty$$

An important detail about regression algorithms is that they must *construct* their estimates, meaning the function type of $\psi_N(\mathcal{S})$ must be determined beforehand. If we let $\mathcal{F}_N$ be the class of functions containing all possible $\psi_N(\mathcal{S})$ and assume it is

[7] By Lusin's theorem ([SS05], Theorem 4.5) and the measurability of f (and by restrictions on f we will make later), f is continuous on sets of large measure, and so we should expect the value of f at a point to usually approximate f on a neighborhood of positive probability.

[8] The ε allows for logarithmic terms in the convergence rate.

a Hilbert subspace of $L^2(\phi) = \{f : \int_{\mathbb{R}^D} f(\mathbf{x})^2 \phi(\mathbf{x}) d\mathbf{x} < \infty\}$, we can break (1.1) apart even more:

Remark 1. *If $\mathcal{F}_N$ is a Hilbert space with respect to the $L^2(\phi)$ norm, then there is a function $f_N = \arg\min_{f_N \in \mathcal{F}_N} \mathbb{E}\left[(f_N - f)^2\right]$, and moreover by the Pythagorean theorem,*

$$\left\|\hat{f}_N - Y\right\|_{MSE}^2 = \mathbb{E}\left[(\hat{f}_N - f_N)^2\right] + \mathbb{E}\left[(f_N - f)^2\right] + \mathbb{E}\left[(f - Y)^2\right]$$

The first two terms are called the estimation error *and* approximation error*, respectively ([DGL96], Chapter 12).*

This factorization reveals a tradeoff in choosing $\mathcal{F}_N$: we want it to be large enough that the approximation error tends to zero, i.e. any function in $L^2(\phi)$ can be approximated by a function in $\mathcal{F}_N$ for large enough N; however we also need it to be small enough that we are able to find an $\hat{f}_N$ that approximates f_N. A consistent regression algorithm thus needs to pick from a space that can approximate an arbitrary regression function well, and needs to be able to effectively leverage sample information into an approximation. We will say $\mathcal{F}_N$ is "approximable" if

$$\lim_{N \to \infty} \inf_{f_N \in \mathcal{F}_N} \mathbb{E}\left[(f_N - f)^2\right] = 0$$

Note the difference between "approximability" and "consistency": approximability is a property of a family of function classes, that it is eventually dense in $L^2(\phi)$; consistency is a property of a regression algorithm, that it can in expectation produce asymptotically perfect predictions of an arbitrary function in $L^2(\phi)$. Some examples of approximable $\mathcal{F}_N$ are as follows:

- Polynomials, with degree increasing in N

  Polynomial regression is the first type of regression taught, with degree-zero polynomial regression taught to elementary schoolchildren as "taking the mean". Undergraduates then learn about linear regression and regression using higher-degree polynomials in a linear algebra class via the Ordinary Least Squares

algorithm. Polynomials are a convenient class to work with, especially since many real-world phenomena can be described with them[9]. The Law of Large Numbers ([ER09], Theorem 4.2.1) is our most basic result regarding the consistency of estimating a constant f (although usually given in terms of the max norm, rather than L^2):

Fact 1 (The Weak Law of Large Numbers)**.** *If f is constant, then the sample mean $\frac{1}{N}\sum_{i=1}^{N} y_i$ is a consistent estimator of f.*

For non-constant estimates, we have approximability by the Stone-Weierstrass theorem ([Rud76], Theorem 7.30)

Fact 2 (Stone-Weierstrass Theorem)**.** *Any continuous function on a compact domain can be approximated with a polynomial.*

Since by Lusin's theorem ([SS05], Theorem 4.5) continuous functions with compact support are dense in $L^2(\phi)$, this also implies universal approximation on $L^2(\phi)$.

- Splines, with resolution increasing in N

  The problem with polynomial approximations is that the number of terms required to approximate well can grow quite fast, and thus also the difficulty in finding the best estimate. In the same vein, polynomials can be too expressive, as with a large enough degree one can fit the observed samples exactly. We could instead reticulate our domain into small compact subsets, for example into grids, and stitch together approximations on each of these regions. These reticulated splines then allow us to have a global function composed of local approximations. Splinal functions are a common sight in approximations in introductory measure

[9] Philosophically, this is probably putting the cart before the horse, as we likely wouldn't care as much about polynomials if they *weren't* useful.

theory courses ([SS05], Theorem 4.1), as measurable functions are defined as a limit of simple functions, which in turn are merely splines of constant functions.

Fact 3 (Measure Theory)**.** *Any continuous function on a compact domain can be approximated with simple functions.*

Indeed, beyond approximability, constant splines are a consistent estimator ([GKKW02], Theorem 4.1):

Fact 4. *Assuming the partitioning of the splines becomes finer as N increases and satisfies certain properties, the constant splinal estimate of the empirical average over each partition is a consistent estimator.*

- Neural networks, with width or depth increasing in N

  Neural networks can be thought of as iterated splines—they consist of a sequence of compatible affine transformations A_i with w_i many rows, along with a univariate *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$. Abusing notation to let σ_i represent the function $\mathbb{R}^{w_{i+1}} \to \mathbb{R}^{w_{i+1}}$ described by stacking[10] the σ, the network yields

  $$\hat{f} = \sigma_\ell \circ A_\ell \circ \cdots \circ \sigma_1 \circ A_1$$

  When the activation function is the (popular) Rectified Linear Unit (ReLU)

  $$\sigma_i^j(x) = \max(0, x + b_{i,j})$$

  it zeros out a half-plane corresponding to $\langle \mathbf{x}, \mathbf{a}_{i,j} \rangle \leq -b_{i,j}$, where the affine transformation A_i is written

  $$A_i(\mathbf{x}) = \begin{pmatrix} \mathbf{a}_{i,1} \\ \vdots \\ \mathbf{a}_{i,w_i} \end{pmatrix} \mathbf{x} + \begin{pmatrix} b_{i,1} \\ \vdots \\ b_{i,w_i} \end{pmatrix},$$

  making the interpretation as iterated splines more clear.

[10] In practice this activation function can also vary across layers, or even across nodes in a single layer. The activation function could also be parameterized, with its parameter learned along with the affine transformations.

Fact 5 (Universal Approximation Theorems)**.** *Any continuous function can be approximated arbitrarily well with a neural network with depth* $\ell = 1$ *and unbounded width* $\max w_i$ *([GKKW02], Theorem 16.1) or with width* $\max w_i \leq D + 4$ *and unbounded depth* ℓ *([LPW+17], Theorem 1).*

Moreover, [GKKW02] Theorem 16.1 gives consistency of wide, shallow, network estimators.

While approximability is a nice property to have, it is in some sense the least we could ask of a regression algorithm: that its domain be capable of providing as small an error as possible. Our concern is thus in how well we can *estimate* the f_N given a sample $\mathcal{S}$. Typically, we would like to say that the function $\hat{f}$ that minimizes the empirical risk

$$\frac{1}{N} \sum_{\mathbf{x},y \in \mathcal{S}} (\hat{f}(\mathbf{x}) - y)^2$$

is close to the regression function f. How effective this strategy is depends on the algorithm, and one could modify this to incorporate other known properties of $\mathcal{F}_N$ or the specific problem, complicating theoretical analysis.

Instead, remember the difficulty of working with finite samples, in that our data could have come from an entire coset of f given by the memorization offsets $m^{\delta}_{\mathcal{S}}$ and any scalar multiple thereof. Since our regression algorithm must return the same result for samples drawn from any $f + \lambda m^{\delta}_{\mathcal{S}}$ (after all, each of those functions yields the same data), and those samples can be arbitrarily different on a set of positive measure, the estimation error currently can be arbitrarily large. This problem is displayed in Figure 1-1, in which the lack of assumed control over the regression algorithm (blue line) means that f is allowed to take any value at the red line. Moreover, if we allow $\hat{f}$ to be any of these $m^{\delta}_{\mathcal{S}}$ cosets, we have a "falsifiability" issue: no data could favor one hypothesized λ multiplier over another[11].

[11] This is similar to the "No Free Lunch" theorems in optimization ([WM97], Theorems 1 and

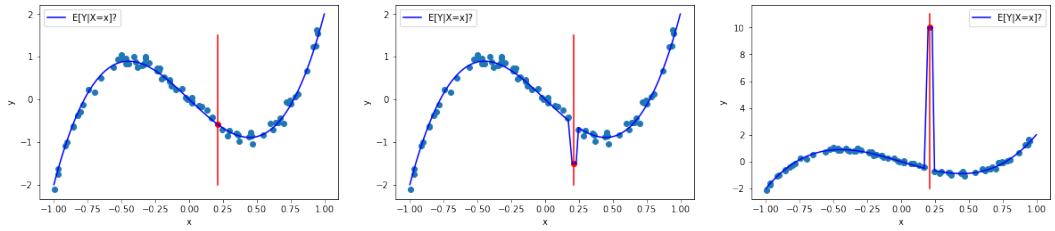

Figure 1-1. Potential graphs of $f(\mathbf{x}) = \mathbb{E}[Y \,|\mathbf{X} = \mathbf{x}]$ (blue line) consistent with observed data (blue dots).

Notice that there is no control over what value $f(\mathbf{x})$ takes at the red line.

We could assert that the regression function belongs to one of the classes we have already defined—for example that it is literally a polynomial—or that it has some other special form, such as in logistic regression where $f(\mathbf{x}) = (1 + \exp(\mathbf{x} \cdot \nu))^{-1}$. The benefit here is that the classes are typically sufficiently restrictive so that it is very difficult to have a falsifiability problem: if two logistic curves are equal (or even close) at sufficiently many points, the curves overall cannot be too different. Additionally, these classes are usually *parameterized*, meaning that there is a map from $\mathbb{R}^k$ into the class that is smooth with respect to MSE[12] and so some optimization algorithm, like gradient descent, can be used to find the minimum.

On the other hand, restricting to such a parametric model is a very strong assumption to make. We will instead make a *non-parametric* assumption, that the function f satisfies certain smoothness properties ([GKKW02], Definition 3.3) to restrict the behavior of $m_{\mathcal{S}}^{\delta}$:

Definition 3. *A function $f : \Omega \to \mathbb{R}$ is (p, C) smooth if f has $k = \lfloor p \rfloor$ derivatives*[13] *and for any multi-index $\mathbf{a}$ with $|\mathbf{a}| = k$,*

$$|\partial^{\mathbf{a}} f(\mathbf{x}) - \partial^{\mathbf{a}} f(\mathbf{x}')| \leq C \|\mathbf{x} - \mathbf{x}'\|^{p-k}$$

2), in that without any knowledge of how your observations relate to unseen datapoints, making better-than-random predictions is impossible.

[12]This smoothness statement is important, as for example the set of functions $\mathbb{Q} \to \mathbb{Q}$ has the same cardinality as $\mathbb{R}$ and thus there always exists a map between them.

[13]Here $\lfloor x \rfloor$ is the largest integer *strictly* less than x; a $(1, 5)$-smooth function is 5-Lipschitz continuous, but not necessarily continuously differentiable.

In this case we say $f \in \mathcal{C}^p(C)$.

By appropriately choosing the behavior in the $\delta/2$-δ-neighborhood of the dataset $\mathcal{S}$ (for example through mollifiers), $m_{\mathcal{S}}^{\delta}$ can be made to have arbitrarily many derivatives, however as p increases or δ decreases, $m_{\mathcal{S}}^{\delta}$ must be scaled down to comply with the uniform bound C on its derivatives. This scaling reduces—and importantly, bounds—the amount of variation within the coset, allowing for algorithm-agnostic discussion of the estimation error. The effect of this restriction on our example function is shown in Figure 1-2, where we assume $(1, 10)$-smoothness[14].

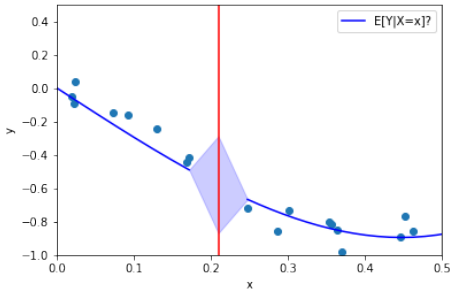

Figure 1-2. Effect of the assumption that $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is $(1, 10)$-smooth, with possible values f could take shaded in light blue.

The question remains, then, of how much of a problem are these memorization cosets given the (p, C)-smoothness assumption? One cause for concern is the reliance on distance, due to the infamous Curse of Dimensionality ([GKKW02], Chapter 2.2):

Fact 6 (Curse of Dimensionality)**.** *It requires a number of datapoints exponential in* D *to achieve a certain density of points. In particular, if* X_i, $i = 1, \ldots, N+1$, *are i.i.d. uniformly distributed in* $[0, 1]^D$, *then the* L^∞ *distance between* $\mathbf{X}_{N+1}$ *and its nearest neighbor is expected to be quite large, as*

$$\mathbb{E}\left[\min_{i=1,\ldots,N} \max_{j=1,\ldots,D} \left|X_{N+1}^j - X_i^j\right|\right] \geq \frac{D}{2D+1} N^{-1/D}$$

[14] It is also common for functions to be described as (k, α)-smooth, $p = k + \alpha$, leaving the constant unstated. This is not the notation that will be used in this document.

One may also view the Curse of Dimensionality in terms of Euclidean distance, in which the volume of the sphere of radius $\frac{1}{2}$ has volume $\approx \left(\frac{\pi e}{2D}\right)^{D/2}$, which is a super-exponentially-small fraction of the volume of the unit cube (volume $1^D = 1$) it inscribes. Assuming points are uniformly distributed, this implies an exponentially large number of points will be required to make it likely that a random point will be within a given radius of any given point.

Indeed, let us assume for the time being that $\mathbf{X}$ is uniformly distributed on $[0,1]^D$. Because our data is so spread out, we can take δ to be rather large, in which case there is plenty of room for $m_{\mathcal{S}}^{\delta}$ to smoothly change from 0 to 1 with minimal need for scaling to compensate. In fact, dimensionality has a major role in how well any regression algorithm can do.

As a concrete example, imagine we observed $N = (h+1)^D$ datapoints equally spaced on a grid in $[0,1]^D$ of side length h^{-1}, with $y_i \equiv 0$ for all i, and that there is no noise: $y_i = f(\mathbf{x}_i)$. Consider two candidate functions that could yield this data:

$$f_1(\mathbf{x}) = 0, \quad f_2(\mathbf{x}) = (h\pi)^{-p} \prod_{j=1}^{D} \sin(h\pi x^j)$$

Both f_1 and f_2 are $(p,1)$-smooth, and yet the L^2 distance between them is $\frac{1}{2}(h\pi)^{-2p}$. Since $h \approx N^{1/D}$, no regression algorithm could guarantee a convergence rate faster than $N^{-\frac{p}{D}}$. Indeed, a similar result holds up to logarithmic factors when the data are assumed to be distributed randomly in the unit cube, rather than precisely at gridpoints ([BDK+17], Theorem 2)[15].

The convergence rate is even slower in the case of noise in the Y values. Assume $Y - f(\mathbf{X})$ is distributed according to a standard normal distribution and is independent of $\mathbf{X}$. We then have a lower bound[16] for the convergence rate of any regression algorithm

[15]This result bounds the supremum norm of the error, which in turn bounds the mean square error. It is however not directly comparable to the forthcoming lower bound in the case of noise, as this theorem bounds the rate of $\liminf_{N\to\infty} \inf_{\{\hat{f}_N\}} \sup_{f\in\mathcal{C}^p(C)}$, which allows the "worst-case" f to change as the number of observations grows and is not reflective of our mindset of "how does error decrease as more observations of a given function are available."

[16]The use of $m_{\mathcal{S}}^{\delta}$ before was for illustrative purposes and is a little too basic to be used in a proof;

([GKKW02], Theorem 3.3)

Fact 7 (Non-parametric Minimax Lower Bound)**.** *Let b_N be an arbitrary positive sequence tending to zero. Then*

$$\inf_{\left\{\hat{f}_N\right\}} \sup_{f \in \mathcal{C}^p(C)} \limsup_{N \to \infty} \frac{\mathbb{E}\left[\left\|\hat{f}_N - f\right\|_{MSE}^2\right]}{b_N N^{-\frac{2p}{2p+D}}} > 0$$

where the infimum is taken over all sequences of estimates $\left\{\hat{f}_N\right\}$ and the supremum over all (p, C)-smooth functions f.

The ordering of quantifiers in this bound states that for any regression algorithm—represented here as a sequence of estimates $\left\{\hat{f}_N\right\}$—there exists a (p, C)-smooth function for which the MSE does not decay asymptotically faster than $N^{-\frac{2p}{2p+D}}$.

It is important to recognize that this is a fundamental issue with non-parametric regression—it does not matter how well-designed a neural network is, or how much computing power is used to produce the estimates; without further assumptions on the nature of the function being regressed it is *not possible* to guarantee being able to distinguish functions below this cursed rate[17]. The issue is that, due to the Curse, there will be many regions of appreciable probability that only contain a single observation $(\mathbf{x}_i, y_i)$—it is then likely that some of these regions will have uncharacteristically large noise, confounding our regression estimate to an even larger degree than was seen in the noiseless case. To further exemplify the problem, we can write the sample complexity by solving for N, letting $\varepsilon^2 = \mathbb{E}\left[(f_1 - f_2)^2\right]$:

$$N \approx \varepsilon^{-\frac{D}{p}-2}$$

The sample complexity is *exponential* in D, a real problem in the modern world where megapixel images exist and sensors are small enough to fit several hundred on a

it might be the case that f has a derivative exactly equal to C at several points, in which case the possible multiples of m_S^δ would be heavily constrained.

[17]It may, however, be the case that the functions f that occur in real world problems are also more likely to be produced by a certain regression algorithm, effectively reducing the problem to a parametric one *by accident*. One may also subvert this rate by swapping the order of the sup and inf statements and tailoring a regression algorithm to a given task.

device. There are then two ways of avoiding the curse: increasing the smoothness p or decreasing the dimension D. Increasing the smoothness such that $p \in \mathcal{O}(D)$ would remove the exponential sample complexity, however this potentially results in the coefficient of the convergence rating being exponentially large in D. Consider the derivatives of $f(x) = \sin(2x)$:

$$f^{(n)}(x) = 2^n \left. \frac{d^n}{dt^n} \sin(t) \right|_{t=2x}$$

As the number of derivatives increases, the bound on their norm must increase exponentially. and with the potential exponential growth in the magnitude of the derivative, we cannot simply appeal to higher smoothness to avoid the curse.

The exponential dependence on dimension should not be surprising. As schoolchildren, we learn the scientific method and the importance of running experiments where only one factor varies from the control group. However, if we have upwards of 42 different binary factors we wish to investigate, the stack of papers describing the 2^{42} different experiments we would need to do would reach the Moon! Instead, we benefit heavily from our innate ability to know what factors are likely extraneous without experiment: the quality of a paper airplane likely doesn't depend on the day of the week it's made, for example. You don't need millions of pixels to identify that a picture is of a cat, as seen in Figure 1-3. For this reason, we would like to supplement our regression algorithms with ways to identify what features are extraneous and what features are meaningful.

In this document, we will continue with a description of various methods of dimension reduction in Chapter 2. Following that, we introduce our algorithm, Multiplicatively Perturbed Least Squares, in Chapter 3. We then provide various experiments showing the practical effectiveness of the algorithm in Chapter 4. Finally, in Chapter 5, we prove the $N^{-1/2}$ consistency of MPLS.

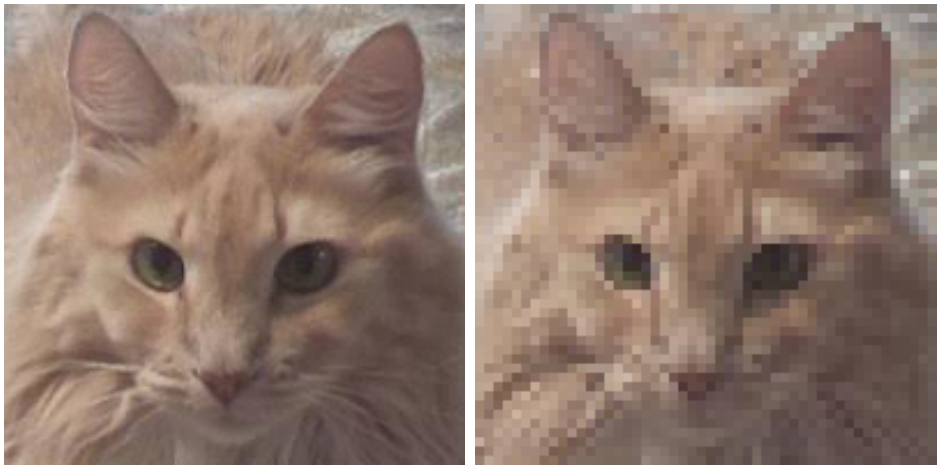

Figure 1-3. Image of a cat, left 200x200 pixels, right 50x50 pixels.
Note how despite being a sixteenth the size, the right image is still clearly identifiable as a cat.

Chapter 2

Dimension Reduction

"A rose by any other name would smell as sweet" - Juliet, on sufficient dimension reduction in botany.

The problem we face is that of sufficient linear dimension reduction; finding a d-dimensional subspace Φ of the ambient $\mathbb{R}^D$ that contains the most information about Y. In [PD09], a projection A onto Φ is a sufficient dimension reduction if one of three conditions holds:

- Inverse reduction: $\mathbf{X}|(Y, A\mathbf{X}) \sim \mathbf{X}|A\mathbf{X}$, signifying that Y does not yield additional information about $\mathbf{X}$ beyond knowledge of $A\mathbf{X}$.
- Forward reduction: $Y|\mathbf{X} \sim Y|A\mathbf{X}$, signifying that Y is not dependent on qualities of $\mathbf{X}$ outside of $A\mathbf{X}$.
- Joint reduction: $\mathbf{X} \perp\!\!\!\perp Y|A\mathbf{X}$, signifying that neither $\mathbf{X}$ nor Y inform the other beyond what is known about $A\mathbf{X}$.

We will assume that the regression function[1] $f : \mathbb{R}^D \to \mathbb{R}$ factors through this projection, in that there exists a $d \times D$ matrix A and a function $g : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(\mathbf{x}) = g(A\mathbf{x})$$

[1]Note that we are free to expand the domain of f, previously denoted as Ω, as the previous definitions rely only on the support of the probability density

This g is often called a *link function*. Note, however, that g and A are not unique—given any invertible $d \times d$ matrix R, we can define a function $g' : \mathbb{R}^d \to \mathbb{R}$ by $g'(\mathbf{z}) = g(R^{-1}\mathbf{z})$, and thus

$$f(\mathbf{x}) = g'(RA\mathbf{x})$$

Importantly from the perspective of the quantitative non-parametric smoothness assumption (Definition 3), if R is not an orthogonal matrix then g' will have derivatives with magnitudes that differ from those of g. To standardize this, we will assume that A is an orthogonal projection, by taking the QR decomposition[2] of $A^T = QR$, where Q is $D \times d$ with orthogonal columns and R an invertible $d \times d$ matrix[3] and letting $g^*(z) = g(R^T\mathbf{z})$. In this case,

$$f(\mathbf{x}) = g((QR)^T\mathbf{x}) = g^*(Q^T\mathbf{x})$$

We will thus assume that A is an orthogonal projection, from which we may also talk about $A^\perp = I_D - A^T A$. We will also often conflate the matrices A and $A^\perp$ with the vector spaces defined by their images, Φ and $\Phi^\perp$ respectively.

For full sufficient dimension reduction, we also need to assume that the noise term is independent of $\mathbf{X}$ conditioned[4] on $A\mathbf{X}$. There remains the issue of the dimension reduction space Φ not necessarily being unique: for example, the rate of global warming depends equally well on the concentration of CO_2 in the atmosphere as it does on the concentration of *everything but* CO_2. Similarly, if two features were independently noisy realizations of an unknown, useful, feature, for example taking measurements of altitude and temperature instead of air pressure, either feature could form a valid dimension reduction space. The difficulty in these examples comes from dependence between intrinsic (CO_2, air pressure) and ambient (non-CO_2 gases, altitude, temperature) measurements. Because of this, dimension reduction algorithms often enforce restrictions on the probability density of $\mathbf{X}$.

[2] Or rather, the RQ decomposition of A.

[3] It will also be upper triangular and have positive diagonal entries, but that isn't relevant here.

[4] We will be assuming that the noise is fully independent of $\mathbf{X}$.

We will make a strong assumption on the distribution of the $\mathbf{X}$, namely that $A\mathbf{X}$ and $A^{\perp}\mathbf{X}$ are independent (which we will call A-independence). This assumption ensures that the multiplicatively-perturbed least squares solutions asymptotically belong to the low dimensional space, i.e. that

$$\mathbb{E}\left[w(\mathbf{X})Y(A^{\perp}\mathbf{X})\right] = 0.$$

In ordinary least-squares, when $w \equiv 1$, mean independence[5] between $A\mathbf{X}$ and $A^{\perp}\mathbf{X}$ suffices, which often arises through an ellipticity assumption on ϕ. As noted in Lemma 14, the multiplicative perturbation we use will be approximately constant, and so even without the A-independence assumption one expects the found solutions will lie close to A; Experiment 4.6 shows that this holds in practice. From the perspective of $N^{-1/2}$ consistency, however, this is an unacceptable (i.e. asymptotically nonzero) source of error.

The class of distributions which satisfy the A-independence requirement contains most distributions which are used to test dimension reduction algorithms, which are usually normal distributions or uniform distributions over axis-aligned rectangles. For clarity, here are some examples and non-examples of distributions in $\mathbb{R}^5 = \mathbb{R}^2 \times \mathbb{R}^3$ for which the projection A onto the first two coordinates is independent from the other three coordinates:

- The standard isotropic multivariate Gaussian is A-independent
- An elliptical multivariate Gaussian where e_1 and e_2 are not singular vectors is not A-independent
- The uniform distribution over $[0, 1]^5$ is A-independent
- The uniform distribution over a rotated cube is in general not A-independent

[5] $\mathbb{E}\left[A\mathbf{X} \,\middle|\, A^{\perp}\mathbf{X}\right] = \mathbb{E}\left[A\mathbf{X}\right]$

- The uniform distribution on the unit ball centered at the origin is not A-independent
- The product distribution of any distribution on $\mathbb{R}^2$ and any distribution on $\mathbb{R}^3$ is A-independent

Interestingly, in Experiment 4.6 we test MPLS's performance when the projection is not aligned with the axes, and still achieve the $N^{-1/2}$ convergence rate. Future work will investigate why this beneficial phenomenon occurs.

Additionally, while dimension reduction only attempts to find "a" basis for the intrinsic space, there is certainly something to be said for finding a "correct" basis, especially when attempting to write down an analytic formula for the function. As noted in [BGM08], unless the building blocks of the function are rotation-invariant, a suboptimal basis can hinder attempts to find a formula. Consider the following functions $\mathbb{R}^2 \to \mathbb{R}$ that are equal up to a rotation and scaling in the domain:

$$f_1(x, y) = \sin(x) + y^3$$

$$f_2(u, v) = u^3 - 3u^2v + 3uv^2 + \sin(u)\cos(v) + \cos(u)\sin(v) - v^3$$

However, in this thesis we will not discuss attempts to find a preferred basis—the assumption will be that the downstream regression algorithm will not benefit substantially from this, for example k-nearest neighbors or polynomial splines.

There have been numerous algorithms for dimension reduction developed over the past 30 years. The general strategy of these algorithms is to attempt to find directions in $\mathbb{R}^D$ where the function value either varies quickly (indicating a direction likely to live in A) or slowly (indicating a direction likely to be orthogonal to A). In Table 2-I, we compare the statistical complexity, computational complexity, and theoretical assumptions of various dimension reduction algorithms. In the following sections, we provide a description of the general philosophies of these methods.

Algorithm	Stat. clx.	Comp. clx.	Assumptions	Remarks
GCR [LL20]	$N^{-1/2}D\exp\left(\sqrt{D}\right)$	N^3D	$A^\perp\mathbf{X}$ elliptically distributed	See Appendix A for further statistical complexity analysis.
MAVE [XTLZ02]	$\mathcal{O}\left(\left(\frac{\log N}{N}\right)^{\frac{3r}{D+4}}\right)$	N^2D^{2r}	The density of $\mathbf{X}$ has bounded fourth derivative and all moments exist.	r is the degree of the polynomial approximation used.
MPLS	$(\log N)N^{-1/2}D^{3/2}$	ND^2	$A\mathbf{X}$ and $A^\perp\mathbf{X}$ independent and sub-Gaussian	No differentiability of f required, however the identified space may have dimension $< d$ (but larger than PHD).
OPG [XTLZ02]	$\mathcal{O}\left(\left(\frac{\log N}{N}\right)^{\frac{2r}{D+4}}\right)$	N^2D^{2r}	The density of $\mathbf{X}$ has bounded fourth derivative and all moments exist.	r is the degree of the polynomial approximation used.
PHD [Li92]	$N^{-\frac{1}{2}}\sqrt{D-d}^\dagger$	ND^2	$\mathbf{X}$ elliptically symmetric	Interpretation as mean Hessian requires f twice differentiable. Space identified may have dimension $< d$.
SAMM [DJS08]	$(\log N)\,N^{-\frac{2}{3\vee d}}$	$N^2D^2\log N$	Positive density on a high probability open set	$\sqrt{N}$ consistent only when $d \leq 4$.
SAVE [LZ07]	$\mathcal{O}(N^{-1/2})^\dagger$	$ND^2 + D^3$	$\mathbf{X}$ is elliptically distributed	Requires a bias correction to the base algorithm to be consistent.
SIR [Li91]	$N^{-1/2}\sqrt{D}^\dagger$	$D(N + D^2)$	$\mathbf{X}$ is elliptically symmetric	Fails when one component is symmetric.

Table 2-I. Comparison of various dimension reduction algorithms. Big-$\mathcal{O}$ notation indicates that there may be suppressed constants that depend on D.
†: Asymptotic result (via Central Limit Theorem)

2.1 Vector Search

The most basic method would be to discretize the unit sphere in D dimensions and search for directions along which the function does not vary much. In [JLT09], a similar method is described for a different style of problem, in which the function is convolved against an oriented kernel in each direction, recording only the directions in which the convolution is small. This, however, has the unfortunate effect of having cursed computational complexity—the size of an ε-net on the D-sphere is exponential in D. In cases where data is scarce relative to computational power this tradeoff may be acceptable.

Alternatively, instead of looking at all directions on the unit sphere, one may look at all $\binom{N}{2}$ directions implied by the data. Contour regression, from [LZC05] and further refined in [LL20], considers pairs of points where either the difference in corresponding Y values is small (in Simple Contour Regression (SCR)) or where the variance of Y values corresponding to a tube around the two points is small (in Generalized Contour Regression (GCR)). The directions then correspond to vectors spanning the orthogonal complement $A^{\perp}$. However, SCR suffers from non-monotonicity in f, and while the change to variance for GCR ameliorates that issue, theoretical understanding of the algorithm (Appendix A) suggests an exponential dependence on D in the constants of the convergence rate, overshadowing the achievement of the fast rate. Additionally, the N^3 computational complexity[6] is rather slow compared to other algorithms.

2.2 Unsupervised Learning

It is, of course, impossible *a priori* to learn the dimension reduction space without using the Y labels. However, it is not unreasonable to expect that—or at least to check—if there is certain geometry to the distribution $\mathbf{X}$, then this might imply properties of

[6] N^3 in the improved version from [LL20], the original in [LZC05] is N^4

the central mean subspace. For example, if our data lies on a hypercube, then one might expect that A is aligned with the faces of the cube. Once the orientation of the cube is determined, the search space of directions is reduced from exponential in D ($\mathcal{O}(\varepsilon^{-D})$ for the ε net on the sphere) to polynomial ($\mathcal{O}\left(\binom{D}{d}\right)$ for the number of d-dimensional sub-bases). Unfortunately, actually determining the orientation of a cube in high-dimensional space is a poorly understood task, and looking for more general heterogeneities is even more difficult. In Appendix B, we provide an algorithm that can find the orientation of a cube and compare the performance of this unsupervised method to more standard supervised algorithms, including MPLS. As shown in the experiments, unsupervised learning methods are not bound by the minimax curse of dimensionality and can potentially achieve much faster convergence rates for identifying the intrinsic subspace. For purposes of downstream regression, however, this benefit does not impact the overall error rate—the $N^{-\frac{2p}{2p+d}}$ will still dominate.

2.3 Inverse Regression

The first algorithm for finding the dimension reduction space in general was Sliced Inverse Regression (SIR) in [Li91]. SIR was the first of a class of algorithms that looked at level sets of the function in order to determine the active subspace. The key insight is that, for centered data, the vector

$$\mathbb{E}\left[\mathbf{X}\,|Y\right]$$

always lives in A; the focus on $\mathbb{E}\left[\mathbf{X}\,|Y\right]$, compared to the regression function $\mathbb{E}\left[Y\,|\mathbf{X}\right]$ derives the name "inverse regression." Thus, by looking at the mean vectors for several level sets one can find a basis for a subspace of the dimension reduction space. If one is lucky, this might reveal the full space, however there are broad classes of functions for which SIR fails to identify A: even functions (with respect to the probability density)

and specifically radial functions will always yield the same mean vector. Symmetry in just one variable will prevent identification of that component, and while identification of some aspect of the subspace may be useful, one needs to identify the entire subspace in order to fully avoid the dependence on D in downstream tasks.

Two additional sliced methods have been introduced since 1991: Sliced Average Variance Estimation (SAVE) from [Coo00] and Smallest Vector Regression (SVR) from [LMV20]. Whereas SIR computes the mean of various level sets to determine the dimension reduction space, SAVE computes the global covariance of the level sets, similarly expecting that the primary variance in the distribution of the level sets will lie in A. As noted in [LZ07], these covariances then must be corrected to account for a nonzero bias in their estimation in order to achieve consistency. Moreover, it appears that SAVE is more sensitive to the number of slices chosen than SIR is.

SVR, on the other hand, is a single-index method ($d = 1$) in which the subspace is estimated for each level set as the singular vector corresponding to the smallest singular value of the centered level sets. Here, the intuition is that level sets of sufficiently monotone functions should be slices of the domain that are narrow in precisely the direction of the single index. This method is difficult to generalize to higher intrinsic dimensions ($d > 1$), unfortunately, as the geometry of level sets in higher dimensions is much more complicated to characterize[7] and the condition of "sufficiently monotone" is harder to justify.

Sliced methods also have a minor disadvantage in that regression problems with discrete ranges—for example binary classification—require special treatment, as one has less control over what slices can be made and the concentration of samples in each slice. If one considers binary $\{0, 1\}$ classification to be a regression problem on the Bayes' classifier $\mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$ with noise $\varepsilon(\mathbf{X})$ taking values $1 - \mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$

[7] This is the classic stumbling block in introductory topology, that compact connected subsets of $\mathbb{R}$ are all intervals, yet compact connected subsets of $\mathbb{R}^n$ for $n > 1$ are not exclusively products of intervals.

and $-\mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$, then there are only two slices available for sliced methods to work with.

2.4 Iterative Kernel Least-Squares Estimates

The Vector Search and Inverse Regression methods can be thought of as looking at the behavior of global linear and constant (respectively) approximations to the data, with the assumption that points that are approximated well should have markedly different structure between the intrinsic and extrinsic dimensions. Other algorithms reverse this approach, relying on the fact that the gradient of the function necessarily lives in A via the Chain Rule

$$\nabla [g(A\mathbf{x})] = \nabla g(A\mathbf{x})A$$

Thus, through computing local linear approximations at various points of the function, one acquires a tranche of vectors primarily living in a low dimensional vector space, a basis of which can then be estimated by taking the top singular vectors of those approximations. The complicating factor is determining the local linear approximation: by the Curse, exponentially few samples live within δ of a given point. For this reason, several methods use some form of Kernel Least-Squares estimation, in defining a kernel function K and minimizing[8] for each j

$$\frac{1}{N}\sum_{i=1}^{N} K(B_t^T(\mathbf{x}_i - \mathbf{x}_j))(y_i - (a_j + b_j^T B_t^T(\mathbf{x}_i - \mathbf{x}_j)))^2 \qquad \text{MAVE}$$

$$\frac{1}{N}\sum_{i=1}^{N} K(B_t^T(\mathbf{x}_i - \mathbf{x}_j))(y_i - (a_j + b_j^T B_t^T(\mathbf{x}_i - \mathbf{x}_j)))^2 \qquad \text{OPG}$$

$$\frac{1}{N}\sum_{i=1}^{N} K_t\left(\left\|B_t^T(\mathbf{x}_i - \mathbf{x}_j)\right\|\right)(y_i - (a_j + b_j^T(\mathbf{x}_i - \mathbf{x}_j)))^2 \qquad \text{SAMM}$$

In all of Minimum Average Variance Estimation (MAVE) from [XTLZ02], Outer Product of Gradients (OPG) from [XTLZ02], and Structural Adaptation via Maximum

[8] in MAVE, B_t and the a_j, b_j are jointly minimized in each iteration; in OPG they are minimized through alternating least squares.

Minimization (SAMM) from [DJS08], the weights are determined by a kernel function applied to the norm of the projected difference between $\mathbf{x}_i$ and $\mathbf{x}_j$—in MAVE and OPG, this kernel is a symmetric differentiable probability density, while in SAMM the kernel is compactly supported. In both cases, this proceeds in an iterative fashion, with the projection matrix being updated each iteration as well as the "bandwidth" of the kernel being narrowed. It is worth noting, however, that in order to achieve $\sqrt{N}$ consistency MAVE and OPG require a polynomial approximation of higher order than linear[9]. It appears that a polynomial approximation of degree $\approx D$ is required for the theory to yield $\sqrt{N}$ consistency, however degree D polynomials over $\mathbb{R}^D$ have $\mathcal{O}(e^D)$ terms, making this workaround cursed in computational time. SAMM, on the other hand, only has $\sqrt{N}$ consistency when $d \leq 4$ (with $N^{\frac{2}{d}}$-consistency otherwise).

2.5 Principal Hessian Directions

Similar to how the gradient of a low-dimensional function lies in the low-dimensional space, the Hessian $H_f(\mathbf{x})$ of a function is a low-rank matrix whose columns span a subspace of the low dimensional space. This has the benefit that the average Hessian, $\mathbb{E}[H_f(\mathbf{X})]$ can potentially span the low dimensional space, and so slicing is not necessary. Moreover, if the data is normally distributed, Stein's Lemma ([Ste81], Lemma 2) states that, with R being the residual of an affine approximation to Y and $\mathbf{X}$ centered,

$$\mathbb{E}[H_f(\mathbf{X})] = \mathbb{E}\left[R\mathbf{X}\mathbf{X}^T\right]$$

[9] Theorem 1 of [XTLZ02] states that the subspace error is $\mathcal{O}(h^3 + h^{-1}\delta_n^2)$, where

$$\delta_n^2 = \frac{\log n}{nh^D}$$

$$\lim_{n\to\infty} \frac{nh^D}{\log n} = \infty$$

The limit statement requires that $h \in \omega\left(\left(\frac{\log n}{n}\right)^{1/D}\right)$, and so the subspace error is $\mathcal{O}\left(\left(\frac{\log n}{n}\right)^{3/D}\right)$, which is cursed.

In the algorithm Principal Hessian Directions (PHD) [Li92], this result is used to estimate the intrinsic space, by taking the top d principal components of this matrix.

2.6 Multiplicatively Perturbed Least Squares

MPLS offers a re-formulation of the PHD algorithm in terms of least-squares linear approximations to the data. A more detailed description of the algorithm will be given in Chapter 3, however the main similarity is through its reliance on Assumption 5, that the vectors $\mathbb{E}\left[R\mathbf{X}\mathbf{X}^T\right]\mathbf{z}_i$, for various $\mathbf{z}_i$, approximately span the low-dimensional space. We also expand the assumptions from strict normality to sub-Gaussianity of the distribution of $\mathbf{X}$ and also do not require any differentiability conditions on Y. The formulation of the estimate in terms of slopes of linear approximations also allows for better incorporation of the initial linear approximation into the final estimate, whereas in PHD the relationship between the mean Hessian matrix and the slope of the linear approximation is less clear.

2.7 The Issue of Computational Complexity

Some of these algorithms have computational complexities that are somewhat slow, such as the N^3 rate in GCR and the N^2 rates of SAMM and MAVE. These are not, *a priori*, problematic—there is a real difference between computational and statistical complexity, and it is easy to believe that for many real world problems acquiring more computational time is cheaper than acquiring more data, if the latter is even possible. However, as noted in the Vector Search section, given a large enough computational budget, one can forgo using one of these fancy dimension reduction algorithms and instead brute-force search an ε net of the unit sphere to find directions in which $\mathbb{E}\left[Y\,|\langle\mathbf{X},\nu\rangle\right]$ is small, which has complexity on the order of $\varepsilon^{-D}ND$.

This is particularly an issue when evaluating these algorithms on toy problems.

Many of these algorithms are published with experimental results on problems where $D = 10$ and $N \approx 1000$. In these cases, a computational complexity of N^2D^2 is about 10^8, which would allow for $\varepsilon \approx 0.4$ in the brute-force case. The N^3D complexity of GCR provides $\varepsilon \approx 0.25$. These are reasonably small errors from the brute-force algorithm, and so experiments with small D may overstate performance simply due to luck happening to allow the algorithm to find a good projection. This is especially of concern in the case of iterative algorithms like MAVE and OPG; the non-trivial chance of finding a good projection randomly, given the computational cost, might explain why it performs better than its theoretical bound in Experiment 4.7.

Chapter 3

Multiplicatively Perturbed Least Squares

Multiplicatively Perturbed Least Squares (MPLS) is a computationally efficient, adaptive, and simple algorithm for finding the dimension reduction space. Philosophically, it falls between kernel methods, like OPG and SAMM, and the second moment method PHD. Similar to the former, it computes local linear approximations to a function at select points, however instead of solving a Kernel Least Squares problem, it solves Ordinary Least Squares on a multiplicative perturbation. These perturbations allow for measurement of a quantity similar to the average Hessian matrix measured by PHD, however the flexibility of picking test points and the benefit of some third moment information yields benefits in some problems.

3.1 Main Results and Discussion

MPLS estimates the low-dimensional subspace as the top right singular vectors of a matrix of "slope perturbations", in the following way:

Algorithm 1 (MPLS)**.** *MPLS has two parameters, M and k. M should be taken to be at least $d \log d$ (to satisfy Assumption 5) and $k \approx D^{-1}$ (to satisfy Assumption 6). The data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is assumed (Assumption 2) to be drawn from a distribution satisfying $\mathbb{E}[\mathbf{X}] = 0$. Pick M points $\mathbf{z}_1, \ldots, \mathbf{z}_M$ in $\mathbb{R}^D$, for example a random subset*

of the $\mathbf{x}_i$'s.

1. *Compute a least squares affine approximation to the data $\{(\mathbf{x}_i, y_i)\}$,*

$$\hat{\beta}, \hat{b} := \underset{b\in\mathbb{R},\beta\in\mathbb{R}^D}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} (y_i - \langle \beta, \mathbf{x}_i\rangle - b)^2$$

2. *Compute the residuals of this approximation, $r_i = y_i - \left\langle \hat{\beta}, \mathbf{x}_i \right\rangle - \hat{b}$*

3. *For each $\mathbf{z}_m$, center the residuals to their weighted mean,*

$$\tilde{r}_{m,i} = r_i - \left(\frac{1}{N}\sum_{i=1}^{N} \exp\left(-k\left\|\mathbf{x}_i - \mathbf{z}_m\right\|^2\right)\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \exp\left(-k\left\|\mathbf{x}_i - \mathbf{z}_m\right\|^2\right) r_i$$

and compute the slope perturbation as the solution to the least squares problem:

$$\hat{\mathbf{p}}_m := \underset{\hat{\mathbf{p}}\in\mathbb{R}^D}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} \left(\exp\left(-k\left\|\mathbf{x}_i - \mathbf{z}_m\right\|^2\right)\tilde{r}_{m,i} - \langle \hat{\mathbf{p}}, \mathbf{x}_i\rangle\right)^2 \tag{3.1}$$

4. *Let $\hat{P}$ be the $M \times D$ matrix whose rows are the $\hat{\mathbf{p}}_m$'s, $m = 1, \ldots, M$, and compute the rank-d singular value decomposition of $\hat{P} \approx U_d \Sigma_d V_d^T$. Define $\hat{A} := V_d^T \in \mathbb{R}^{d\times D}$.*

There are two parameters that need to be chosen in MPLS: k and M. The parameter M serves as the estimate of the intrinsic dimension, however the guess need not be tight at all: in the experiments of Chapter 4, M is taken to be $20d$. We note, however, that this does impact the computational complexity—the cost of uncertainty of d is additional runtime, as the computational complexity is linear in M. The estimate $\hat{A}$ is derived from the top singular vectors of $\hat{P}$; without prior knowledge of d, one could instead use the singular values to determine d. As shown in the proof of Theorem 1, the singular values corresponding to directions in A should be on the order of $\frac{\log N}{D}$, while those in $A^\perp$ will be $\sqrt{\frac{D(\log N)^2}{N}}$.

In Chapter 5, we prove the $N^{-1/2}$ consistency of this algorithm (Theorem 1) with some modifications to allow for easier theoretical analysis, namely:

1. Assuming $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$ and $\mathbb{E}[Y] = 0$, and thus defining $\hat{\beta}$ as the slope of a linear approximation to the data.

2. Partitioning the data into 2 subsets, computing $\hat{\beta}$ separately on one of them and computing the slope perturbations $\hat{\mathbf{p}}_m$ from the residuals to β on the remaining data.

A version of MPLS that is faithful to these assumptions is not very different from the algorithm presented here. Satisfying the first modification can be done by estimating $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]$ and $\mathbb{E}[Y]$ on an independent sample and preprocessing the remaining data, while the second is a straightforward modification to the algorithm. However, we expect the theoretical results to be approximately correct even for the unmodified MPLS presented in Algorithm 1.

The full theoretical analysis is left to Chapter 5, however we include an informal description of the theorem here.

Theorem 1 (Informal)**.** *If*

1. *There is a $d \times D$ matrix A with orthonormal rows, a (p, C_f)-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $p \leq 1$ and a vector β such that*

$$\mathbb{E}[Y \,|\mathbf{X} = \mathbf{x}] = f(A\mathbf{x}) + \langle \beta, \mathbf{x} \rangle$$

   and $\mathbb{E}[f(A\mathbf{X})\mathbf{X}] = 0$, and $\mathbb{E}[f(A\mathbf{X})] = 0$, and the noise $Y - \mathbb{E}[Y \,|\mathbf{X}]$ is independent of $\mathbf{X}$ (Assumption 1) and sub-Gaussian (Assumption 4)

2. *The regressors $\mathbf{x}_i$ are independent samples of a sub-Gaussian random vector $\mathbf{X}$, which has the property that $A\mathbf{X} \perp\!\!\!\perp A^{\perp}\mathbf{X}$ (Assumption 3). Without loss of generality, $\mathbb{E}[\mathbf{X}] = 0$, $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$ (Assumption 2).*

3. *The $\mathbf{z}_m$'s are taken to be so that $\|\mathbf{z}_m\| \in \mathcal{O}(\sqrt{D})$ and Assumption 5 is satisfied,*

4. *The parameter k is taken to be $\mathcal{O}(D^{-1})$ satisfying Assumption 6,*

5. *The previous modifications to the MPLS algorithm are made, partitioning the data into* 2 *subsets to estimate* $\hat{\beta}$ *and the* $\hat{\mathbf{p}}_m$ *independently;*

then, with probability close to 1*, exponentially in* N*, there exist constants* C_1*,* C_2*,* C_3*,* C_4*, and* τ_ε*, such that, for all* $t > \tau_\varepsilon(\log_2 2N)^{-1/2}$ *the following two results hold:*

1. *With probability* $1 - 10\exp(-t)$*, if* $N > \max(D^3, 32)$ *then the angle between* A *and the subspace* $\hat{A}$ *identified by the modified MPLS algorithm is bounded:*

$$\left\|\sin\Theta(\hat{A}, A)\right\| \leq \sqrt{\frac{t^2 C_1 D^3 (\log_2(2N))^2}{N}} \left(1 + C_2\sqrt{t}\right)$$

2. *With probability* $1 - 8\exp(-t)$*,*

$$\left\|\beta - \hat{\beta}\right\| \leq t\sqrt{\frac{C_4 D \log_2(2N)}{N}}$$

The constants C_1*,* C_2*,* C_3*,* C_4 *and* τ_ε *may be chosen to depend only on the sub-Gaussian norms of* $\mathbf{X}$ *and* $\mathbb{E}[Y|\mathbf{X}]$*, the smoothness of* f*,* $\log M$*, and the properties of the function* $\mathbf{q}$ *described in Assumption 5.*

This is proven in Chapter 5 as Theorem 1.

Remark 2. *The computational complexity of MPLS is as follows, assuming solving a least squares problem is done via solving the normal equations and matrix inversion of a* $D \times D$ *matrix is* $\mathcal{O}(D^3)$

Step	*Complexity*
Computation of $\hat{\beta}$	$ND^2 + D^3$
Computation of residuals	ND
Pre-computing inverse covariance matrix	$ND^2 + D^3$
Computing weights (M *times)*	MND
Computing $\hat{\mathbf{p}}_m$ *(*M *times)*	MND
Final SVD	ND^2

Thus, the total computational complexity is $\mathcal{O}(ND(M+D)+D^3)$*. With the requirement that* $N > D^3$*, this simplifies to* $\mathcal{O}(ND(M+D))$*.*

One benefit of solving a multiplicatively perturbed least squares problem, as opposed to a weighted least squares problem, is that the inverse matrix used to solve the normal equations is the same for all $\mathbf{z}_m$'s, and thus only needs to be computed once. Computing all the $\hat{\mathbf{p}}_m$ is hence only a matrix multiplication of the pre-computed inverse covariance matrix and the multiplicative perturbations, yielding MND complexity, rather than $M(ND^2 + D^3 + ND)$.

3.2 Comparison with Principal Hessian Directions

Via Assumption 5, the main mechanism behind MPLS is the same as that of Principal Hessian Directions, with a few differences. First is that while PHD requires the mean Hessian $\mathbb{E}\left[f(A\mathbf{X})\mathbf{X}\mathbf{X}^T\right]$ be full rank, we also consider a third moment component $\mathbb{E}\left[f(A\mathbf{X})\left\|\mathbf{X}\right\|^2\mathbf{X}\right]$ that can potentially either complement a rank-deficient Hessian or reinforce directions with small eigenvalues.

We also appreciate the novelty of using Stein's lemma to identify the expectation $\mathbb{E}\left[f(A\mathbf{X})\mathbf{X}\mathbf{X}^T\right]$ as the mean of the Hessian in the case when $\mathbf{X}$ is normally distributed. In MPLS, too, Stein's lemma yields insights into the individual vector solutions that are found. Stein's lemma uses integration by parts on $\mathbb{E}\left[f(A\mathbf{X})\mathbf{X}\right]$ to place a derivative on f by moving the $\mathbf{X}$ term into the Gaussian probability density $\exp\left(-\left\|\Sigma^{-1/2}\mathbf{x}\right\|^2\right)$. Under the A-independence assumption (Assumption 3), the individual slope perturbations (3.1) found by MPLS are of the form

$$\mathbb{E}\left[\exp\left(-k\left\|A(\mathbf{X}-\mathbf{z})\right\|^2\right)\tilde{f}_{m,k}(A\mathbf{X})A\mathbf{X}\right].$$

Notably, because we use a Gaussian weight, we can integrate against the Gaussian weight instead of the probability density function. For example, if the density ϕ is differentiable with respect to the Lebesgue measure on a compact domain Ω with piecewise continuous boundary, then

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-k\left\|\mathbf{X}-\mathbf{z}\right\|^2\right)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right] &= \int \exp\left(-k\left\|\mathbf{X}-\mathbf{z}\right\|^2\right)\tilde{f}_{m,k}\mathbf{X}\phi(\mathbf{X})d\mathbf{X} \\
&= \int \exp\left(-k\left\|\mathbf{X}-\mathbf{z}\right\|^2\right)\tilde{f}_{m,k}(A\mathbf{X})(\mathbf{X}-\mathbf{z})\phi(\mathbf{X})d\mathbf{X} \\
&= \int_{\partial\Omega} -(2k)^{-1}\exp\left(-k\left\|\mathbf{X}-\mathbf{z}\right\|^2\right)\tilde{f}_{m,k}\phi(\mathbf{X})dS \\
&\quad + (2k)^{-1}\int_{\Omega}\exp\left(-k\left\|\mathbf{X}-\mathbf{z}\right\|^2\right) \\
&\quad \cdot\left(\frac{\partial}{\partial\mathbf{X}}\left[\tilde{f}_{m,k}\right]\phi(\mathbf{X}) + \tilde{f}_{m,k}\frac{\partial}{\partial\mathbf{X}}\left[\phi(\mathbf{X})\right]\right)d\mathbf{X}
\end{aligned}$$

In particular, if ϕ is a uniform distribution, its derivative on the interior of its support is zero, and so we get Stein's result plus a boundary term.

Chapter 4

Experimental Results

In this chapter, we examine our algorithm's performance on different problems and compare it to other dimension reduction algorithms: GCR, OPG, PHD, SAVE, SCR, and SIR. We wish to see how, in practice, the following impact the results found:

1. Small (40) vs large (200) D, in Example 4.1
2. Small (1%) vs large (100%) noise, in Example 4.2
3. Oscillations vs Periodicity, in Example 4.3
4. Small (1) vs large (4) d, in Example 4.4
5. Truncated normal vs uniform distribution of $\mathbf{X}$, in Example 4.5
6. Gaussian vs uniform noise, in Example 4.6
7. Comparison of other algorithms in a small $D = 10$ setting, in Example 4.7

In each experiment, noise will be added with a variance given in terms of σ^2, which will be the inherent variance of the function values for that example. Typically, the noise level will be 5%, meaning the variance of the noise will be $1/20^{\text{th}}$ the variance of the function, i.e. its squared L^2 norm[1]. We do this to allow for more natural

[1]Specifically, the squared L^2 norm of $f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]$

comparisons across examples; a fixed variance noise has a different effect on the behavior of a high-variance function than on that of a low-variance one; the fixed variance would mean different effective variances for regressing $f(\mathbf{x})$ and $2f(\mathbf{x})$.

It is perhaps worth mentioning that many of these tests will be much harder than those in the analysis of previous literature: in [Li91] (SIR), [Li92] (PHD), [XTLZ02] (MAVE), [LZC05] (GCR), [LL20] (GCR), the experiments all set $D = 10$ at most, with some experiments taking D as small as 4. The default set-up for our experiments will be $D = 40$, which because of the curse of dimensionality results in a much harder problem. We do this because of the concern raised in Appendix A, which is that there may be cursed constants in the convergence rates; a convergence rate of $N^{-1/2}D\exp\left(\sqrt{D}\right)$, as demonstrated in the bound of GCR in Appendix A, is not theoretically useful for reasonable values of N, as it requires $N \geq D^2\exp\left(2\sqrt{D}\right)$ many samples for this fraction to drop below 1, which will be very large for even moderately large D. When $D = 10$, this quantity is about 55,000, which is a large but not unreasonable[2] value for N; when $D = 40$, it is about $5 \cdot 10^8$. By choosing $D = 40$, then, we hope to better demonstrate the efficiency of MPLS and existing algorithms in higher dimensions. This does, however, result in our experiments having larger N than the previously-cited literature, in which typically $N \approx 500$; here we will typically take $N \leq 3{,}200$, with certain experiments testing N up to 51,200.

In each example, we will report the sine of the angle between the regressed subspace $\hat{A}$ and the true subspace A, that is

$$\sin\theta = \left\|\hat{A}\hat{A}^T - AA^T\right\|_2.$$

While reporting the true angle, rather than its sine, allows for greater granularity in the $\theta \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right]$ range[3], this is not the region of interest to us. Moreover, for the purposes of theory the sine is more relevant, both in that our theoretical analysis

[2]The experiments in [LL20] test N up to $10^{4.5} \approx 32{,}000$.

[3]In other words, if an algorithm is succeeding in reducing the angle from $\frac{\pi}{2}$ to $\frac{\pi}{4}$, that change is more significant on a plot of the angle ($1.57 \to 0.785$) than on a plot of the sine ($1 \to 0.707$)

(Theorem 1) bounds the sine and in the utility of considering the norm $\left\|\hat{A}\hat{A}^T - AA^T\right\|_2$, rather than its arcsin.

The reported error, $\sin\theta$, is then the mean of 100 random starts, with error bars indicating the t-test 95% confidence interval. The results are plotted on a $\log_{10} - \log_{10}$ plot, on which power relations appear as straight lines. Achieving the $N^{-1/2}$ convergence rate would be realized as a line with slope $-1/2$.

For each set of parameters (N, noise level, random start #) within an Example, each algorithm in the suite is run on the same $(\mathbf{X}, Y)$ data. Moreover, while we will describe the low-dimensional functions as depending only on the first d components of the input, the data given to the algorithms is actually QX, where Q is a random $D \times D$ orthogonal matrix, and thus the rows of A are the first d columns[4] of Q. This is to avoid the issue raised in Chapter 2.1 and examined in Appendix B, that the problem of dimension reduction becomes trivial if we know the orientation of the axes; here, the central mean subspace is almost surely not aligned with the coordinate axes.

The parameters of MPLS are chosen as follows: the weight parameter k is chosen so that the median weight is $1/4$ in all examples[5] i.e.

$$k = \frac{\log 4}{\text{median } \|\mathbf{x}_i - \mathbf{z}\|^2}$$

The $\mathbf{z}_m$'s are chosen uniformly at random among the datapoints, with $M = 20d$. Additionally, as MPLS determines both a global linearity of the function and a basis for the intrinsic subspace, we must decide how to combine these two pieces of information to produce our regressed subspace. We choose one of two methods, depending on the experiment:

- Hybridized:

  For some experiments, we add the global unweighted solution to the matrix of

[4] Except in Experiment 4.6, in which the rows of A are linear combinations of columns of Q

[5] Recall from Lemma 13 that a k that satisfies Assumption 6 results in a empirical mean weight of at least $1/4$. It is, however, easier to solve for k given a desired median weight than a desired mean weight, hence this method.

slope perturbations row-wise, i.e.

$$\tilde{P} = \hat{P} + M^{-1/2}\mathbb{1}\hat{\beta}^T$$

The $M^{-1/2}$ factor has experimentally been useful to ensure the noise of the $\hat{\beta}$ estimate does not overshadow the slope perturbations. The estimated subspace is then the space spanned by the top singular vectors of $\tilde{P}$.

- Purely non-linear:

  In other experiments, we disregard $\hat{\beta}$ entirely and report the subspace spanned by the top singular vectors of $\hat{P}$.

We will indicate which choice is made in each experiment; the choice is made purely based on which option performs better. In practice, one could run options and compare their performance on a validation set. The same choice is made for PHD, where in the first case $D^{-1/2}$ times $\hat{\beta}\hat{\beta}^T$ is added to the mean Hessian matrix.

The code for SAVE and SIR was used from the Python package *sliced* ([Loy18]) with default parameters, while Wenjing Liao and Hao Liu generously provided Matlab code for SCR, GCR, and SIR (from their analysis in [LL20]). We ran both the Python and Matlab version of SIR for these experiments with default parameters, unsurprisingly with indistinguishable results; we report the results of the Python version. I wrote the code for OPG and PHD myself.

Due to the super-linear computational complexity of GCR (N^3) and OPG (N^2), they are excluded from most experiments; Experiment 4.7 is done with smaller D and N and includes these algorithms.

4.1 Example 1: L2SqL4 and the Effect of D

Let $\mathbf{X} \sim \mathcal{N}(0, I_D)$, where D will be 40 or 200, $d = 4$, with the function f_1, which we will call "L2SqL4" as it is a difference of an L_2 norm and a squared-L_4 norm, defined

as

$$f_1(x, y, z, w) = \sqrt{(x+1)^2 + (y-1)^2} - \sqrt{(z-1)^4 + (w-1)^4} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.05\sigma^2)$ where σ^2 is the variance of the f_1 values, and let $N = 400$, 1600, 3200, 6400, 12800, 25600, 51200. Here we hybridize the linear and slope perturbation estimates in MPLS.

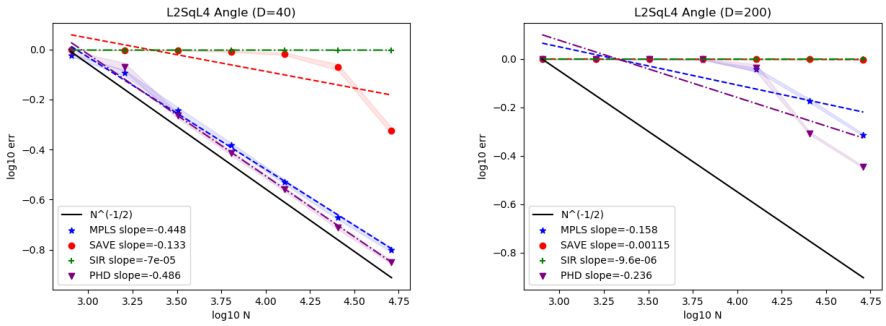

Figure 4-1. Angle of regressed subspace for L2SqL4, left $D = 40$, right $D = 200$.

In this example, we the expected theoretical performance from MPLS, achieving the $N^{-1/2}$ convergence rate almost immediately when $D = 40$ and appearing to require a certain number of samples before achieving the rate when $D = 200$. Interestingly, neither SIR nor SAVE appear to find any foothold in regressing A, although SAVE perhaps would start to perform better with a few more samples in the $D = 40$ case. It is not surprising that SIR fails[6], as L2SqL4 likely has many symmetries that prevent SIR from identifying the subspace. SAVE likely similarly suffers from similar-looking level sets, although it appears to find just enough data to latch onto in the $D = 40$ case. Additionally, the minimum requirements for N implied by the theory for MPLS can be seen in the $D = 200$ plot, as the error only begins to drop once N reaches a critical value, at which point the convergence rate appears to approximate $N^{-1/2}$. It is interesting, however, that the minimum requirements of Theorem 1 appear to be a large overestimate in this case, as the requirement $N \geq D^3$ would imply $N \geq 8 \cdot 10^6$

[6]Unfortunately for SIR, most of our experiments are done with symmetric functions. Experiment 4.5 will show a successful performance by SIR.

for $D = 200$, which is much larger than 51,200. Even when $D = 40$, the minimum theoretical requirement is $N \geq 64,000$, which is larger than any sample in these experiments.

4.2 Example 2: Dalalyan3 and the Effect of Noise

Let $\mathbf{X} \in [0, 20]^{40}$ be uniformly distributed, $d = 3$, with the function f_2, which we shall call "Dalalyan3" as it is the third example function used in [DJS08], used there for the purpose of examining different noise levels and defined as

$$f_2(x, y, z) = (1 + x)(1 + y)(1 + z) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ will have $\sigma^2 = 1\%$ or 100% of the variance of the f values, and let $N = 200, 400, 800, 1600, 3200$. Here we hybridize the linear and slope perturbation estimates in MPLS.

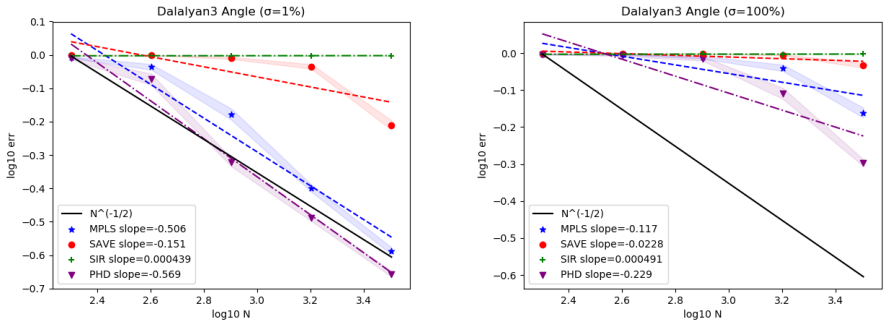

Figure 4-2. Angle of regressed subspace for Dalalyan3, left $\sigma^2 = 1\%$, right $\sigma^2 = 100\%$.

As can be seen in the plots, there is a considerable deterioration in behavior for all algorithms in the high-noise scenario. Both MPLS and PHD appear to require a certain threshold of N before achieving the $N^{-1/2}$ rate[7].

[7]This behavior is expected from the theoretical results of Theorem 1, in which the constant τ_ε depends on the variance of the noise and is a component of many of the constants

4.3 Example 3: Ripple vs Radial Cosine and the Effect of Oscillations vs Periodicity

Let $\mathbf{X} \in [-2, 2]^{40}$ be distributed uniformly, $d = 2$, with two functions defined in terms of $r = \|\mathbf{x}\|_2$:

1. f_3, which we shall call "Ripple" as its graph looks like ripples with increasing amplitude

$$f_3(r) = (r^2 + 1)\cos(\pi r) + \varepsilon,$$

2. g_3, which we shall call "Radial Cosine"

$$g_3(r) = \cos(\pi r) + \varepsilon$$

where in both cases $\varepsilon \sim \mathcal{N}(0, 0.05\sigma^2)$, where σ^2 is the variance of the respective function, and $N = 200, 400, 800, 1600, 3200$. Here we discard the linear estimate and use only the slope perturbation estimates in MPLS.

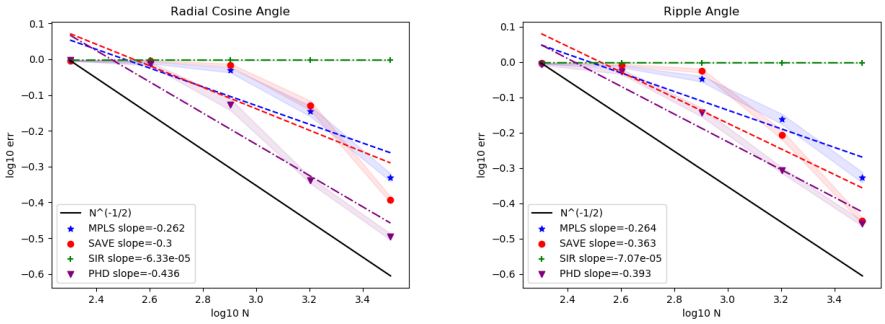

Figure 4-3. Angle of regressed subspace, left Ripple, right Radial Cosine

In these cases, we can see that oscillations are very difficult for subspace regression algorithms to work with as it is not until larger N that MPLS, PHD, and SAVE begin to achieve the $N^{-1/2}$, with MPLS and SAVE appearing to "catch up" to the fast rate as the sample size increases.

4.4 Example 4: L1 and the Effect of d

Let $\mathbf{X} \in [-d^{-1/2}, d^{-1/2}]^{40}$ be distributed uniformly, $d = 1$ or $d = 4$, and define a function f_4, which we shall call "L1" as it is the L_1 norm of its input

$$f_4(\mathbf{x}) = \sum_{i=1}^{d} |x_i| + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 0.05\sigma^2)$ where σ^2 is the variance of f_4 and $N = 800, 1600, 3200, 6400, 12800$. Here we discard the linear estimate and use only the slope perturbation estimates in MPLS.

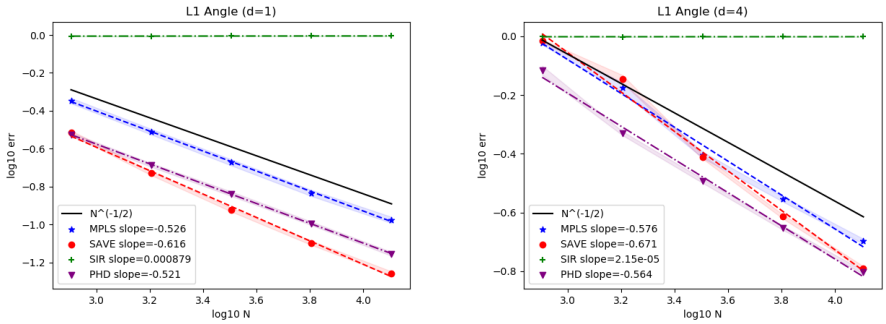

Figure 4-4. Angle of regressed subspace for L1, left $d = 1$, right $d = 4$

The subspace is slightly more difficult to identify for larger d, with the regressed error being about a constant multiple worse in this case. However, all of MPLS, PHD, and SAVE achieve the $N^{-1/2}$ rate in both scenarios.

The domain is artificially shrunk between the two examples to preserve the same overall variance of the outputs of f_4. If we scaled f_4 by $d^{-1/2}$ instead of resizing the domain, or didn't scale at all, we get similar results (Figure 4-5).

4.5 Example 5: Li2 and the Effect of Distribution

Let $\mathbf{X} \in [-2, 2]^{40}$ be distributed either uniformly or according to a truncated standard normal distribution, $d = 2$, and define a function f_5, which we shall call "Li2" as it is

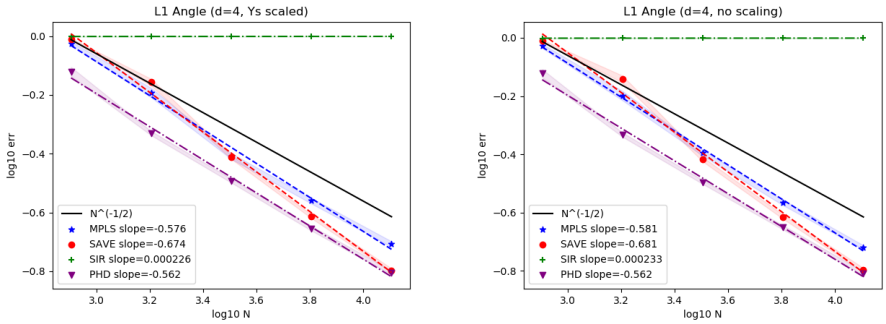

Figure 4-5. Angle of regressed subspace for L1 on $[-1,1]^D$, left f_4 is scaled, right no scaling.

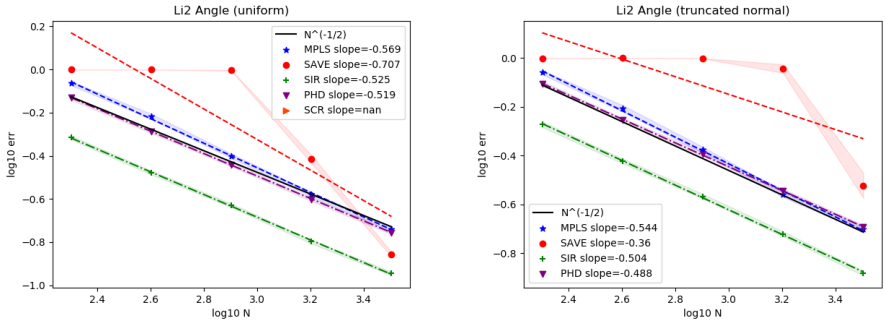

Figure 4-6. Angle of regressed subspace, left uniform, right normal

the second example function from [Li91]:

$$f_5(x,y) = \frac{4x}{2+(2y+3)^2} + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 0.5\sigma^2)$, σ^2 being the variance of f_5, and $N = 200, 400, 800, 1600, 3200$. Here we hybridize the linear and slope perturbation estimates in MPLS.

First, f_5 is not symmetric on this domain, which is shown in that SIR does well on this problem. The distribution appears to only have a minor impact on MPLS, PHD, and SIR; although there is potentially a difference in the behavior of SAVE. In both cases, MPLS, PHD, and SIR all immediately achieve the $N^{-1/2}$ rate, while SAVE appears to require a threshold number of samples before identifying A.

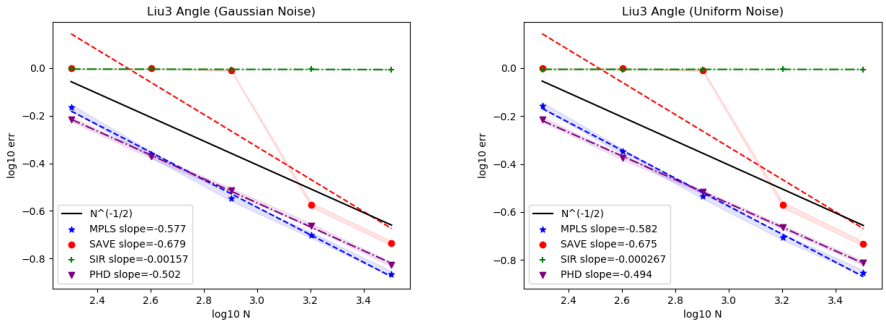

Figure 4-7. Angle of regressed subspace, left Gaussian noise, right uniform noise

4.6 Example 6: Liu3 and the Effect of Noise Distribution

Let $\mathbf{X} \in [-1,1]^{40}$ be distributed uniformly, $d = 2$, and define a function f_6, which we shall call "Liu3" as it is the third example function from [LL20]:

$$f_6(x,y) = \sin\left(\frac{\pi}{2}(x-1)\right) + y + \varepsilon$$

with $\mathbb{E}[\varepsilon^2] = 0.05\sigma^2$ being distributed either as a Gaussian or uniformly, where σ is the standard deviation of f_6, and $N = 200, 400, 800, 1600, 3200$. Here we hybridize the linear and slope perturbation estimates in MPLS.

Additionally, for this function the components are not axis-aligned: $x = \frac{1}{3}\sum_{i=1}^{9} x_i$ and $y = x_{10}$. This does not satisfy our assumption that the intrinsic and ambient spaces are independent, although they are uncorrelated.

Interestingly, the violation of the independence assumption does not appear to have hurt us in this case, as MPLS (and PHD) achieves the $N^{-1/2}$ rate immediately in both cases, even as SAVE and SIR fail to do so[8]. SAVE, however, does succeed in achieving the $N^{-1/2}$ rate for N sufficiently large.

[8] $\sin(x - \frac{\pi}{2}) = -\cos(x)$, which is an even function on this symmetric domain, so it is not suprising that SIR fails.

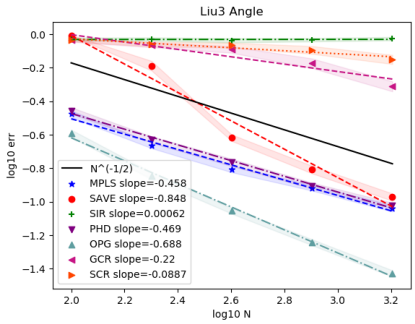

Figure 4-8. Angle of regressed subspace, $D = 10$, additional algorithms reported

4.7 Comparison with Other Algorithms

As previously mentioned, the algorithms OPG, SCR, and GCR all have computational complexities that depend on N super-linearly: N^2 in the case of OPG and SCR and N^3 in GCR. For computational time reasons, we include a smaller example here. We use the same set-up as in Example 4.6, with the difference being that $\mathbf{X}$ is now uniformly distributed in $[-1, 1]^{10}$, and $N = 100, 200, 400, 800, 1600$.

In this experiment, only MPLS, PHD, OPG, and eventually SAVE achieve the $N^{-1/2}$ rate on this range of N's. Interestingly, OPG manages a rate faster than $N^{-1/2}$; this is potentially due to the hidden benefit of increased computational time for small D discussed in Section 2.7. SCR is recognized in [LZC05] to struggle with non-monotonicities, which this function has, and hence GCR does perform better than SCR on this problem. However, it appears that if GCR does achieve the $N^{-1/2}$ rate, it is not until N is significantly larger.

Chapter 5

Main results and their proofs

5.1 Problem set up and assumptions

We state formally our assumptions, then the main result, and then dedicate the rest of this chapter to its proof. Recalling from Chapter 2, our goal is to identify the central mean subspace Φ, which has the property that, letting S_Φ be the orthogonal projection[1] onto Φ,

$$\mathbb{E}\left[Y\,|\mathbf{X}\right] = \mathbb{E}\left[Y\,|S_\Phi\mathbf{X}\right]$$

We consider two subspaces of Φ by decomposing the regression function $\mathbb{E}\left[Y\,|\mathbf{X}\right]$ into a linear function and its residuals, as follows:

Assumption 1. *The regression function $\mathbb{E}\left[Y\,|\mathbf{X}=\mathbf{x}\right]$ is inherently low dimensional, i.e. there is a $d \times D$ matrix A with orthonormal rows, a vector $\beta \in \mathbb{R}^D$, and an (p,C_f)-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $p \leq 1$ such that*

$$Y = f(A\mathbf{x}) + \langle \beta, \mathbf{x} \rangle + \varepsilon \tag{5.1}$$

and

$$\mathbb{E}\left[f(A\mathbf{X})\right] = 0, \quad \mathbb{E}\left[f(A\mathbf{X})\mathbf{X}\right] = 0, \quad \mathbb{E}\left[f(A\mathbf{X})^2\right] = V_f.$$

The noise, ε, is assumed to be independent of $\mathbf{X}$.

[1] The ability for us to assume S_Φ has orthonormal rows without loss of generality was shown in (2).

The space Φ is thus spanned by β and the rows of A, which we estimate independently; let Φ_A denote the row space of A. Because of this, if β is not contained in Φ_A, the dimension of the central mean subspace will be $d+1$—otherwise, it will be d. The $d+1$ case represents the special case of a partial linear model, where the regression function relies on a feature in a purely linear fashion; [WJ03] provides a short history of interest in the partial linear model, albeit from the perspective of identifying β rather than A.

It is also worth noting that the (p, C_f)-smoothness of f does not result in any issues with p being small in our proofs. These parameters are used in Lemma 10 to bound Y on compact subsets in terms of the diameter of these subsets. Because of our definition of (p, C_f)-smoothness in Definition 3 and the property $\mathbb{E}\left[f(A\mathbf{X})\right] = 0$, we have that $|f(A\mathbf{x})|$ is bounded by $2C_f(1 + \|A\mathbf{x}\|)$. This behavior is different from what we expect in regression where the goal is to predict y values based on observations, and hence higher differentiability constrains how much the regression function can change over small distances. Instead, this assumption helps control the behavior of quantities like $Y\mathbf{X}$ by bounding the global variance of Y.

Our goal is to show that, with high probability, the solutions to a multiplicative least squares problem approximately span Φ_A. To simplify the theoretical analysis, we assume the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ are partitioned into 2 sets, $\mathcal{S}, \mathcal{S}'$, each of size $\frac{N}{2}$. Define, thus,

$$\hat{\beta} := \operatorname*{argmin}_{\tilde{\beta}} \frac{2}{N} \sum_{i=1}^{N/2} \left(y_i' - \left\langle \tilde{\beta}, \mathbf{x}_i' \right\rangle\right)^2 \tag{5.2}$$

$$r_i := y_i - \left\langle \hat{\beta}, \mathbf{x}_i \right\rangle \tag{5.3}$$

$$\tilde{r}_{m,i} := r_i - \left(\frac{2}{N} \sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m)\right)^{-1} \frac{2}{N} \sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m) r_i \tag{5.4}$$

$$\hat{\mathbf{p}}_m := \operatorname*{argmin}_{\tilde{\mathbf{p}}} \frac{2}{N} \sum_{i=1}^{N/2} \left(w_k(\mathbf{x}_i; \mathbf{z}_m)\tilde{r}_{m,i} - \langle \tilde{\mathbf{p}}, \mathbf{x}_i \rangle\right)^2 \tag{5.5}$$

Since $\hat{\beta}$ is a consistent[2] estimate of β, we should expect r_i to approximate $f(A\mathbf{x}_i)$; Lemma 3 will give bounds on the difference between the two.

To ease the notation, we will make, without loss of generality, the following assumption:

Assumption 2. *The random variable* $\mathbf{X}$ *is centered,* $\mathbb{E}[\mathbf{X}] = 0$*, and its covariance is the identity matrix,* $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$*.*

Note that this combined with Assumption 1 implies that $\mathbb{E}[Y] = 0$. One could define r_i as the residual to an *affine* approximation to y and allow for nonzero $\mathbb{E}[Y]$, however the bounds of consistency for affine approximations are more complicated than those of linear approximations. In practice, $\mathbb{E}[\mathbf{X}]$ and $\mathbb{E}[Y]$ can be estimated with a subset of the training data and subtracted off from the remainder without ill effect. Similarly, $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]$ can be estimated on an independent sample in order to enforce $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$ on the remainder of the data.

As discussed in Chapter 3, some of these transformations can be omitted in practice: if $\hat{\beta}$ is taken as the slope of an affine transformation, the residuals will be the same as if $\mathbb{E}[Y] = 0$ and $\mathbb{E}[\mathbf{X}] = 0$, however $\mathbb{E}[\mathbf{X}] = 0$ will still need to be enforced for the computation of $\hat{\mathbf{p}}_m$. The assumption that $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$ is also only for simplification of theoretical analysis[3], in particular to satisfy isotropicity for Lemma 8. Similar results hold for $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] \neq I$, with additional dependence on its largest and smallest eigenvalues.

Assumption 3. *The random variable* $\mathbf{X}$ *is sub-Gaussian and* A*-separable, i.e. the density function* ϕ *can be written as the product of two independent densities on the range of* A *and* $A^\perp$ *respectively.*

[2]It is not necessarily, however, an *unbiased* estimate for finite N in this case. This is to say that $\lim_{N\to\infty} \mathbb{E}\left[\left\|\hat{\beta}_N - \beta\right\|\right] = 0$, however $\mathbb{E}\left[\left\|\hat{\beta}_N - \beta\right\|\right]$ may be nonzero.

[3]That is, the estimates $\hat{\mathbf{p}}_m$ are computed via the usual least squares normal equations, which includes estimating and inverting the covariance matrix. The fact that the covariance matrix is the identity is not used in the estimation.

Variable	Description
A	$d \times D$ orthogonal matrix
β	Slope of the linear approximation to the regression function
D	Ambient dimension of $\mathbf{X}$
d	Dimension of the central mean subspace
f	Component of the regression function orthogonal to the space of linear polynomials.
N	Number of datapoints $(\mathbf{x}_i, y_i)$ available
Φ	Central mean subspace of the regression function, dimension d or $d+1$
Φ_A	Subspace of Φ of dimension d, row space of A
ϕ	Sub-Gaussian probability density of $\mathbf{X}$
$\psi_{\mathbf{X}}$	Sub-Gaussian norm of $\mathbf{X}$
ψ_A	Sub-Gaussian norm of $A\mathbf{X}$
ψ_ε	Sub-Gaussian norm of $Y - \mathbb{E}[Y \mid \mathbf{X}]$
V_f	L^2 norm of f, $\mathbb{E}[f(A\mathbf{X})^2]^{1/2}$
$\mathbf{X}$	Random vector in $\mathbb{R}^D$ of observations/regressors; $\mathbb{E}[\mathbf{X}] = 0$
$\mathbf{x}_i, y_i$	i.i.d. samples from $\mathbf{X} \times Y$
Y	Random variable in $\mathbb{R}$ of responses

Table 5-I. List of parameters defining the problem.

Let $\psi_{\mathbf{X}}$ be the sub-Gaussian norm of $\mathbf{X}$, that is

$$\psi_{\mathbf{X}} := \sup_{\|\nu\|=1} \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{\langle \mathbf{X}, \nu \rangle^2}{t^2} \right) \right] \leq 2 \right\} \tag{5.6}$$

Note that ψ is defined by projections of $\mathbf{X}$ onto unit vectors, hence it is reasonable to expect it to remain $\mathcal{O}(1)$ even when D is large. Subsection 5.2.4 will discuss solely properties of $A\mathbf{X}$; for this reason, we further define ψ_A to be the sub-Gaussian norm of $A\mathbf{X}$.

Assumption 4. *The noise $\varepsilon := Y - \mathbb{E}[Y \mid \mathbf{X}]$ is sub-Gaussian and independent of $\mathbf{X}$, with sub-Gaussian norm*

$$\psi_\varepsilon := \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{\varepsilon^2}{t^2} \right) \right] \leq 2 \right\} \tag{5.7}$$

The basis of the intrinsic space will be determined through the manner in which approximations to the function $\|A(\mathbf{x} - \mathbf{z}_m)\|^2 f(A\mathbf{x})$ change as $\mathbf{z}_m$ vary. In order for

Variable	Description	Definition
$\hat{\beta}$	Slope of the linear approximation to the data $(\mathbf{x}'_i, y'_i)$	(5.2)
c	Absolute constant connecting the sub-Gaussian norm to the inequalities in Lemma 5	Lemma 5
C_β	Constant involved in the distribution of $\left\|\hat{\beta} - \beta\right\|$	(C.11)
C_R	Constant involved in the distribution of $\mathbf{u}_m - \mathbf{p}_m$	(5.26)
C_S	Constant involved in the distribution of $\hat{S}$	(C.2)
$\tilde{f}_{m,k}$	Offset version of the function f so that $\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\tilde{f}_{m,k}(\mathbf{X})\right] = 0$	(5.14)
η	Error of $\hat{\mathbf{p}}_m$ compared to $\mathbf{u}_m$	(5.20)
k	Bandwidth parameter of order D^{-1}, satisfies Assumption 6	Assumption 6
κ	Normalized bandwidth parameter $\kappa = kD$	(5.13)
K_A	Maximum norm of the normalized projected test points $d^{-1/2}A\mathbf{z}_m$	Assumption 5
$K_{A\mathbf{X}}$	Constant involved in the bound of $\|A\mathbf{X}\|$	(C.8)
$K_{\hat{\beta}}$	Constant involved in the bound of $\left\|\hat{\beta} - \beta\right\|$	(C.12)
K_R	Constant involved in the bound of $\|r_i\|$	(C.13)
$K_{\mathbf{X}}$	Constant involved in the bound of $\|\mathbf{X}\|$	(C.10)
K_Y	Constant involved in the bound of $\|Y\|$	(C.9)
$K_{\mathbf{z}}$	Maximum norm of the normalized test points $D^{-1/2}\mathbf{z}_m$	Assumption 5
λ_Q	Bound on the smallest singular value of $M^{-1/2}Q$	Assumption 5
M	Number of test points at which we compute approximations	Assumption 5
Q	$M \times d$ matrix with rows $\mathbf{q}(\mathbf{z}_m)$	Assumption 5
$\mathbf{q}(\mathbf{z})$	Slope of the asymptotic linear approximation to the $\mathbf{z}$-based multiplicatively perturbed function	Assumption 5
r_i	Residuals of the linear approximation $\hat{\beta}$ to $(\mathbf{x}'_i, y'_i)$	(5.3)
$\tilde{r}_{m,i}$	Offset r_i so that $\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m)\tilde{r}_{m,i} = 0$	(5.4)
$\mathbf{p}_m$	Asymptotic solution to the MPLS problem	(5.17)
$\hat{\mathbf{p}}_m$	Empirical solution to the MPLS problem	(5.5)
ψ_η	Related to the sub-Gaussian norm of η	(5.21)
$\hat{S}$	Sample covariance matrix of $\mathbf{x}_i$, $\frac{2}{N}\sum_{i=1}^{N/2} \mathbf{x}_i\mathbf{x}_i^T$	Lemma 8
$\mathbf{u}_m$	Quantity close to both $\mathbb{E}[\hat{\mathbf{p}}_m]$ and $\mathbf{p}_m$	(5.20)
$w_k(\mathbf{x}; \mathbf{z})$	Gaussian weight with bandwidth k^{-1}, $\exp\left(-k\|\mathbf{x} - \mathbf{z}\|^2\right)$	(5.9)
$\mathbf{z}_m$	One of M test points in $\mathbb{R}^D$	Assumption 5

Table 5-II. List of constants and variables introduced in the proofs.

MPLS to be successful, we must assume that these approximations identify Φ_A. This is the main restriction on our function, however we expect that it should hold for a wide class of functions.

Assumption 5. *For $\mathbf{z} \in \mathbb{R}^D$, let $\mathbf{q}(\mathbf{z}) \in \mathbb{R}^d$ denote the least squares solution to*

$$\mathbb{E}\left[\left(f(A\mathbf{X})\left\|A(\mathbf{X}-\mathbf{z})\right\|^2 - \langle A\mathbf{X}, \mathbf{q}(\mathbf{z})\rangle\right)^2\right] \tag{5.8}$$

Then, the $M \times d$ matrix Q with rows $\mathbf{q}(\mathbf{z}_m)$ is invertible; let λ_Q satisfy $\|Q^{-1}\|^{-1} \geq \lambda_Q\sqrt{M}$. Further, let $K_{\mathbf{z}} := D^{-1/2}\max_m \|\mathbf{z}_m\|$, and $K_A := d^{-1/2}\max_m \|A\mathbf{z}_m\|$.

Since, in practice, one might pick the $\mathbf{z}_m$ at random from the $\mathbf{x}_i$'s, this is often a statement about the second moment matrix of $\mathbf{q}(\mathbf{Z})$, in this case with $\mathbf{Z}$ having the same distribution as X. From here on, the $\mathbf{z}_m$ will be assumed to be fixed—note that the number of $\mathbf{z}_m$ points M does not change with N. The $\sqrt{M}$ factor in the smallest singular value of Q normalizes the constant λ_Q, for as will be seen later in Theorem 6, it is reasonable to expect the singular values of a random matrix to scale with $\sqrt{M}$.

Assumption 5 is not an unreasonable assumption to make: since polynomial approximations to $f(A\mathbf{X})(A\mathbf{X})$ have no constant or linear factor, it would be rather unlucky for $\|A(\mathbf{X}-\mathbf{z}_m)\|^2 f(A\mathbf{X})(A\mathbf{X})$ to have that property as well. In Figure 5-1, we show the effect when $d = 1$ of this multiplicative perturbation.

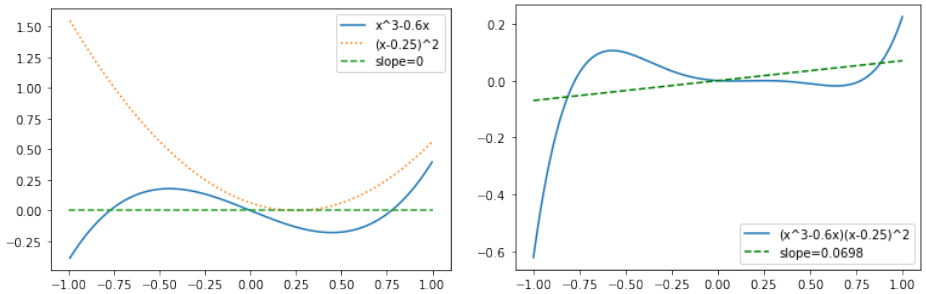

Figure 5-1. Effect of multiplicative perturbation for the function $f(x) = x^3 - \frac{3}{5}x$.

Remark 3. *Note that the* $\mathbf{q}(\mathbf{z}_m)$ *can be expressed, via the normal equations, as*

$$\begin{aligned}\mathbf{q}(\mathbf{z}_m) &= \mathbb{E}\left[(A\mathbf{X})(A\mathbf{X})^T\right]^{-1}\mathbb{E}\left[f(A\mathbf{X})\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 A\mathbf{X}\right] \\ &= \mathbb{E}\left[f(A\mathbf{X})\left\|A\mathbf{X}\right\|^2 A\mathbf{X}\right] - 2\mathbb{E}\left[f(A\mathbf{X})(A\mathbf{X})(A\mathbf{X})^T\right]\mathbf{z}_m\end{aligned}$$

Thus, assumption 5 is equivalent to requiring that

$$\mathbb{E}\left[f(A\mathbf{X})(A\mathbf{X})(A\mathbf{X})^T\right]$$

has rank at least $d-1$ *and* $\mathbf{z}_m$ *span its row space, and, if the rank is* $d-1$*, that* $\mathbb{E}\left[f(A\mathbf{X})\left\|A\mathbf{X}\right\|^2 A\mathbf{X}\right]$ *is linearly independent from the column space of* $\mathbb{E}\left[f(A\mathbf{X})(A\mathbf{X})(A\mathbf{X})^T\right]$.

This remark shows the similarity between the space found by MPLS and that of PHD (see Subsection 2.5): both primarily determine the space $\mathbb{E}\left[f(A\mathbf{X})(A\mathbf{X})(A\mathbf{X})^T\right]$. However we note that the full-rank assumption on $\mathbb{E}\left[f(A\mathbf{X})(A\mathbf{X})(A\mathbf{X})^T\right]$ required for PHD is relaxed to allow, in certain cases, for a one-dimensional null space for this matrix.

These core assumptions are enough to ensure well-behavedness properties of the random variable $f(A\mathbf{X})\mathbf{X}$ and, with high probability, of the sample measurements.

As will be motivated in the next subsection, we will use a Gaussian weight function:

$$w_k(\mathbf{x};\mathbf{z}) = \exp\left(-k\left\|\mathbf{x}-\mathbf{z}\right\|^2\right). \tag{5.9}$$

Despite its similarity to the kernel functions used to localize the least-squares estimate in other algorithms, we will see in Lemma 14 that in our case we will need k to be small, on the order of D^{-1}. The proof of Theorem 4 will also require k to be small in a manner independent of D; we expect that these requirements are weaker than the bound of D^{-1}, and this is certainly the case in the high dimensional regime $D \gg d$ that we are interested in. Therefore, our kernels are not local, having a width dependent on D, and we hence expect enough samples within a constant number of

standard deviations from the center $\mathbf{z}$, independently of D. This is a key factor that makes our construction avoid the curse of dimensionality.

Assumption 6 (Size of k)**.** *Let k be a number such that the following conditions are satisfied for all chosen $\mathbf{z}_m$:*

1. *Relationship to the norm of $\mathbf{z}_m$:*

$$k^{-1} \geq \frac{7}{2} D \max\left(1, K_{\mathbf{z}}^2\right) \tag{5.10}$$

2. *With Q defined in Assumption 5, which only depends on the intrinsic d-dimensional distribution of $\mathbf{q}(\mathbf{Z})$ and $f(A\mathbf{X})$:*

$$k^{-1} \geq \frac{8 \|Q\|_F^2 \left(3 + C_1 d^{5/2}\right)}{3\sigma_d(Q)^2 \min_m \|\mathbf{q}(\mathbf{z}_m)\|} \tag{5.11}$$

   and C_1 is defined in (5.39), only depending on K_Y, K_A, and ψ_A.

3. *For all $\mathbf{z}_m$'s, the following two bounds are satisfied:*

$$k^{-1} \geq \max \begin{cases} 2W_A d^{7/4} \left(\psi_A + K_A\right) \sqrt{8^{1/2} V_f \psi_A} \\ 8\sqrt{2} W_A d^{5/2} V_f \left(\psi_A + K_A\right)^2 \psi_A \end{cases} \tag{5.12}$$

   with W_A defined in Corollary 9.

Define

$$\kappa := kD \tag{5.13}$$

noting that (5.10) implies $\kappa \leq \frac{2}{7}$.

Because we are using the slope of a linear approximation, we want the function to be mean zero under our weights. To do this, we work with offset $f(A\mathbf{X})$ values, $\tilde{f}_{m,k}(A\mathbf{X})$, that have weighted mean equal to zero:

Remark 4. *Define*

$$\tilde{f}_{m,k}(A\mathbf{X}) = f(A\mathbf{X}) - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1} \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m) f(A\mathbf{X})\right], \tag{5.14}$$

so that $\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\right] = 0$.

Let $\overline{f}_{m,k} := \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1} \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m) f(A\mathbf{X})\right]$.

It is worth noting that while $\mathbb{E}\left[\tilde{f}_{m,k}(A\mathbf{X})\right] \neq 0$, the global linear slope is still 0:

$$\mathbb{E}\left[\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right] = \mathbb{E}\left[f(A\mathbf{X})\mathbf{X}\right] - \overline{f}_{m,k}\mathbb{E}\left[\mathbf{X}\right] = 0$$

We do, however, wish to know how large $\mathbb{E}\left[\tilde{f}_{m,k}(A\mathbf{X})\right]$ is; since the introduced weights are rather flat, this should remain small; Lemmas 12 and 13 bound the expected weight and sample mean weight respectively, while Lemma 14 and Corollary 9 bound the variances of the weights and the low-dimensional weights respectively. Thus, via Hölder's inequality, the value of $\mathbb{E}\left[\tilde{f}_{m,k}(A\mathbf{X})\right]$ can be bounded; this is of particular value for Lemmas 1 and 4.

5.2 Statement and Proof of the Main Theorem: $N^{-1/2}$ Consistency of the MPLS Estimate

Theorem 1. *Taking assumptions 1, 2, 3, 4, 5, and 6, let $\mathcal{S}, \mathcal{S}'$ be independent partitions of the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, each of size $\frac{N}{2}$. Define $\hat{\beta}$ to be the minimizer of*

$$\frac{2}{N}\sum_{(\mathbf{x}'_i, y'_i)\in\mathcal{S}'_m}\left(y' - \left\langle\hat{\beta}, \mathbf{x}'\right\rangle\right)^2$$

and let $\mathcal{R} = \left\{\left(\mathbf{x}_i, y_i - \left\langle\hat{\beta}, \mathbf{x}_i\right\rangle\right) : (\mathbf{x}_i, y_i) \in \mathcal{S}\right\}$. Define

$$\tilde{\mathcal{R}}_m = \left\{\left(\mathbf{x}_i, r_i - \frac{\frac{2}{N}\sum_{(\mathbf{x}_j, r_j)\in\mathcal{R}} w_k(\mathbf{x}_j; \mathbf{z}_m) r_j}{\frac{2}{N}\sum_{(\mathbf{x}_j, r_j)\in\mathcal{R}} w_k(\mathbf{x}_j; \mathbf{z}_m)}\right) : (\mathbf{x}_i, r_i) \in \mathcal{R}\right\}$$

Let $\hat{P}$ be the $M \times D$ matrix whose rows $\hat{\mathbf{p}}_m$ are solutions to the least squares problem

$$\hat{\mathbf{p}}_m = \underset{\hat{\mathbf{p}}}{\operatorname{argmin}} \frac{2}{N}\sum_{(\mathbf{x}_i, \tilde{r}_i)\in\tilde{\mathcal{R}}_m}\left(w_k(\mathbf{x}_i; \mathbf{z}_m)\tilde{r}_i - \langle\hat{\mathbf{p}}, \mathbf{x}_i\rangle\right)^2 .$$

Then, there exists constants c and C_S such that with probability at least

$$1 - 2\exp\left(-\frac{cN}{16\psi_{\mathbf{X}}^2 \max\left(C_S^2\psi_{\mathbf{X}}^2 D, 192\right)}\right),$$

there exist values ψ_η (5.21), C_R (5.26), C_β (C.11), τ_ε (C.7) depending on the sub-Gaussian norms of $\mathbf{X}$ and $Y - \mathbb{E}\left[Y\,|\mathbf{X}\right]$ and the smoothness of f such that following two results hold for all $t > \tau_\varepsilon(\log_2 2N)^{-1/2}$:

1. *The subspace $\hat{A}$ spanned by the top d right singular vectors of $\hat{P}$ has an angle with A of order $N^{-1/2}$; more precisely, if $N \geq \max(D^3, 32)$, then with probability greater than $1 - 10\exp(-t)$,*

$$\left\|\sin\Theta(\hat{A}, A)\right\| \leq \sqrt{\frac{t^2 D^3 (\log_2(2N))^3}{N}} \left(\frac{8\left(C_R + \psi_\eta\sqrt{t\log_2 2M}\right)}{\kappa\lambda_Q}\right) \tag{5.15}$$

2. *With probability greater than $1 - 8\exp(-t)$,*

$$\left\|\beta - \hat{\beta}\right\| \leq \sqrt{\frac{t^2 C_\beta D \log_2(2N)}{N}} \tag{5.16}$$

Before giving the proof of this theorem, we now discuss the size and effects of the constants. All constants involved (save for k) should be expected to be independent of D, with some dependence on d and the sub-Gaussian norms of $\mathbf{X}$ and $Y - \mathbb{E}[Y|\mathbf{X}]$.

The parameter M, interestingly, has only a minor impact on the statistical complexity of MPLS. It is not the case that we are hoping to find many approximate "noisy" solutions and have the noise cancel out through the singular value decomposition. Increasing the number of slope perturbations here serves only to amplify both the signal and the noise, resulting in only a logarithmic dependence. We do, however, need $M \geq d$ (to satisfy Assumption 5) in order to actually have a hope of spanning the low-dimensional subspace. Taking $M \approx d\log d$ is sufficient to overcome coupon-collecting problems, and $M = D$ effectively guarantees the $\mathbf{z}_m$ span the low-dimensional space.

Proof of Theorem 1. By Lemma 9, with probability

$$1 - 2\exp\left(-\frac{cN}{16\psi_{\mathbf{X}}^2 \max\left(C_S^2\psi_{\mathbf{X}}^2 D, 192\right)}\right)$$

we have bounds on the eigenvalues of the sample covariance matrices and the norm of the sample means for each partition. The probability bounds of (5.15) and (5.16) are stated as sub-exponential probabilities in a parameter t; let $\tau = \sqrt{t}$, and thus with probability $1 - 6\exp(-t)$ the results of Lemma 10 hold, and with probability

$1-8\exp(-t)$ and the assumption $N > D^3$, the results of both Lemma 10 and Corollary 8 hold[4]. It remains to show that, assuming these results, the bounds on $\left\|\beta - \hat{\beta}\right\|$ and $\left\|\sin\Theta(\hat{A}, A)\right\|$ each hold with probability $1 - 2\exp(-t)$.

Lemma 11 yields immediately the result (5.16) for $\left\|\hat{\beta} - \beta\right\|$.

For the bound on $\left\|\sin\Theta(\hat{A}, A)\right\|$, we show that the matrix of slope perturbations $\hat{P}$ is approximately equal to a matrix P whose columns are a basis of the low dimensional space. Through Wedin's theorem (Theorem 5), we can bound the angle between the span of the top singular vectors of $\hat{P}$ and a low-rank matrix not too different from it.

Our matrix of least squares estimates $\hat{P}$ is comprised of three summands:

$$\hat{P} = P + T_1 + T_2$$

By Wedin's theorem, we have that

$$\left\|\sin\Theta(\hat{P}_d, P_d)\right\| \leq \frac{\|T_1 + T_2\|}{\sigma_d(\hat{P})}$$

- The matrix P is comprised of the asymptotic solutions

$$\mathbf{p}_m = \mathbb{E}\left[\exp\left(-k\left\|\mathbf{X} - \mathbf{z}_m\right\|^2\right)\tilde{f}_{k,m}(A\mathbf{X})A\mathbf{X}\right], \tag{5.17}$$

  which are vectors in Φ_A. By Corollary 3 the smallest nonzero singular value of P is greater than

$$\frac{k\lambda_Q}{4}\sqrt{M}$$

- The matrix T_1 of the differences between the quasi-expected slope perturbations and the asymptotic solutions,

$$\mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2}\tilde{r}_{m,i}\mathbf{x}_i\right] - \mathbf{p}_m\,,$$

  which by Lemma 3 are bounded by

$$C_R t^{3/2}\sqrt{\frac{D(\log_2 2N)^3}{N}}\,,$$

[4]Corollary 8 only requires $N > D^2$, however we take this stronger assumption to match the coefficient of the convergence rate.

and in turn with an additional factor of $\sqrt{M}$ this bounds the largest singular value of T_1.

- The matrix T_2 of the differences between the computed slope perturbations and a vector similar to their expected value[5],

$$\hat{\mathbf{p}}_m - \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2}\tilde{r}_{m,i}\mathbf{x}_i\right],$$

which by Theorem 3 is a sub-Gaussian random vector, and so by Corollary 1, with probability at least $1 - 2\exp(-t)$, the largest singular value of T_2 is bounded by

$$t^2\psi_\eta\sqrt{\frac{DM(\log_2 2N)^3\log_2 2M}{N}}$$

From Weyl's inequality and the respective results, we can bound the smallest singular value of $\hat{P}A$ as

$$\begin{aligned}
\sigma_d(\hat{P}A) &\geq \sigma_d(P) - \sigma_1(T_1A) - \sigma_1(T_2A) \\
&\geq \sigma_d(P) - \sigma_1(T_1) - \sigma_1(T_2) \\
&\geq \frac{k\lambda_Q}{4}\sqrt{M} - C_R\sqrt{\frac{t^3DM(\log_2 2N)^3}{N}} \\
&\quad - t^2\psi_\eta\sqrt{\frac{DM(\log_2(2N))^3\log_2(2M)}{N}} \\
&= D^{-1}\sqrt{M}\left(\frac{\kappa\lambda_Q}{4} - \sqrt{\frac{t^2D^3\log_2(2N)^2}{N}}\left(C_R + \psi_\eta\sqrt{t\log_2(2M)}\right)\right)
\end{aligned}$$

Similarly,

$$\sigma_1(T) \leq \sigma_1(T_1) + \sigma_1(T_2) \leq D^{-1}\sqrt{M}\left(\sqrt{\frac{t^2D^3\log_2(2N)^3}{N}}\left(C_R + \psi_\eta\sqrt{t\log_2(2M)}\right)\right)$$

Thus, via Wedin's theorem we prove our bound. Additionally, since $\|\sin\Theta\| \leq 1$, we use the inequality $\min\left(1, \frac{x}{a-x}\right) \leq \frac{2x}{a}$ for $x < a$ to simplify the result, noting further that $\frac{2x}{a} > 1$ for $x \geq a$. □

[5]This is further divided into $\hat{\mathbf{p}}_m \pm \mathbb{E}[\hat{\mathbf{p}}_m] - \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2}\tilde{r}_{m,i}\mathbf{x}_i\right]$ in the proof

In the following sections, we prove theorems and lemmas that support the main aspects of the proof of Theorem 1; further useful results are proved in Appendix C.

5.2.1 The Least Squares Solution

We start with the result that the solution to a multiplicatively perturbed least squares problem lives in the intrinsic space.

Theorem 2. *Assume* $\mathbf{X}$ *satisfies the distribution and independence conditions of Assumption 3, and that* $w : \mathbb{R}^D \to \mathbb{R}$ *is bounded and A-separable, i.e. there exist functions* $w_A : \mathbb{R}^d \to \mathbb{R}$ *and* $w_\perp : \mathbb{R}^{D-d} \to \mathbb{R}$ *such that*

$$w(\mathbf{x}) = w_A(A\mathbf{x})w_\perp(A^\perp\mathbf{x})\,. \tag{5.18}$$

Let $R = f(A\mathbf{X}) + \overline{R} + \varepsilon$*, where* $\mathbb{E}\left[w_A(A\mathbf{X})R\right] = 0$ *and* ε *is i.i.d. mean zero noise with* $\mathbb{E}\left[\varepsilon\mathbf{X}\right] = 0$ *and* $\mathbb{E}\left[\varepsilon^2\right] < \infty$*. Then, if* $\mathbf{u}$ *minimizes*

$$\mathbb{E}\left[\left(w(\mathbf{X})R - \langle\mathbf{u}, \mathbf{X}\rangle\right)^2\right],$$

then $A^\perp\mathbf{u} = 0$*.*

Proof. The minimum is found via the normal equations; since $\mathbf{X}$ is sub-Gaussian, w is bounded, and ε has finite second moment all expectations exist:

$$\mathbf{u} = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]^{-1}\mathbb{E}\left[w(\mathbf{X})R\mathbf{X}\right]$$

Via the uncorrelation assumption,

$$A^\perp\mathbf{u} = \mathbb{E}\left[(A^\perp\mathbf{X})(A^\perp\mathbf{X})^T\right]^{-1}\mathbb{E}\left[w(\mathbf{X})R\mathbf{X}\right] = \mathbb{E}\left[(A^\perp\mathbf{X})(A^\perp\mathbf{X})^T\right]^{-1}\mathbb{E}\left[w(\mathbf{X})R(A^\perp\mathbf{X})\right]$$

We can then use the independence assumption on $\mathbf{X}$ and the A-separability of w to yield the desired result:

$$\begin{aligned}\mathbb{E}\left[w(\mathbf{X})R(A^\perp\mathbf{X})\right] &= \mathbb{E}\left[w_A(A\mathbf{X})R\mathbb{E}\left[w_\perp(A^\perp\mathbf{X})A^\perp\mathbf{X}\,|A\mathbf{X}\right]\right] \\ &= \mathbb{E}\left[w_A(A\mathbf{X})R\right]\mathbb{E}\left[w_\perp(A^\perp\mathbf{X})A^\perp\mathbf{X}\right] = 0\end{aligned}$$

□

This is a very useful theorem. It says that if we solve the perturbed least squares problem, then the solution lies in A, as

$$\mathbf{u} = \mathbb{E}\left[(A\mathbf{X})(A\mathbf{X})^T\right]^{-1}\mathbb{E}\left[w_\perp(A^\perp\mathbf{X})\right]\mathbb{E}\left[w_A(A\mathbf{X})R(A\mathbf{X})\right] \tag{5.19}$$

However, we have no guarantee that this will occur for finite N, that the solution will have appreciable norm, nor that the solutions (over the various linear perturbations over subsets of samples) will span the space.

There is also the issue of choosing the weight function. The separability requirement (5.18) is rather strong, given that *a priori* we do not know what the subspace A is. One weight function that stands out is the Gaussian:

$$w_k(\mathbf{x};\mathbf{z}) = \exp\left(-k\left\|\mathbf{x}-\mathbf{z}\right\|^2\right) = \exp\left(-k\left\|A(\mathbf{x}-\mathbf{z})\right\|^2\right)\exp\left(-k\left\|A^\perp(\mathbf{x}-\mathbf{z})\right\|^2\right)$$

This has the useful property of being separable across arbitrary subspaces, and so in particular, with some abuse of notation, $w_k(\mathbf{X};\mathbf{z}) = w_k(A\mathbf{X};A\mathbf{z})w_k(A^\perp\mathbf{X};A^\perp\mathbf{z})$.

5.2.2 Estimation Error

In this section we will be bounding the difference between sample computations and their asymptotic values: our goal is the following theorem, which provides the source of the $N^{-1/2}$ convergence rate:

Theorem 3. *Take the assumptions on the distribution of* $\mathbf{X}$ *in Assumptions 2 and 3, the assumptions on the behavior of* Y *in Assumptions 1 and 4, and the assumption on the size of* k *in Assumption 6. The estimation error* η_m *of the least squares solution* $\hat{\mathbf{p}}_m$

$$\hat{\mathbf{p}}_m = \operatorname*{argmin}_{\hat{\mathbf{p}}} \frac{2}{N}\sum_{i=1}^{N/2}\left(w_k(\mathbf{x}_i,\mathbf{z}_m)\tilde{r}_{m,i} - \langle\hat{\mathbf{p}},\mathbf{x}_i\rangle\right)^2,$$

i.e., via (5.19),

$$\eta_m := \hat{\mathbf{p}}_m - \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\tilde{r}_{m,i}\mathbf{x}_i\right] =: \hat{\mathbf{p}}_m - \mathbf{u}_m, \tag{5.20}$$

is sub-Gaussian, with tail bound

$$\mathbb{P}\left[\|\eta_m\| > t\right] \leq 2\exp\left(-\frac{Nt^2}{\tau^6 D(\log_2 2N)^3\psi_\eta^2}\right),$$

where

$$\psi_\eta^2 := 32c^{-1}\psi_{\mathbf{X}}^2\left(75K_R^2 + 2V^2D^{-1}C_S^2\psi_{\mathbf{X}}^2\right) \tag{5.21}$$

where V is defined in (5.23) and C_S in (C.2).

We will bound the sub-Gaussian norm by bounding the tail probabilities of $|\langle\hat{\mathbf{p}}_m - \mathbf{u}_m, \nu\rangle|$ for arbitrary unit vectors ν, via Lemma 5. This will require bounding two separate terms, as demonstrated in the following two lemmata.

Lemma 1. *Taking assumptions 1-4 and 6, and conditioning on the high probability bound on $f(A\mathbf{X})$ in Corollary 8, which entails assuming $N > D^2$, the term $\mathbf{u}_m$ (5.20) cannot be too large, i.e.*

$$\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{r}_m\mathbf{X}\right]\| \leq V\tau^3\sqrt{\frac{(\log_2(2N))^3}{D}} \tag{5.22}$$

where

$$V = \frac{C_R}{\sqrt{\log_2(2M)}} + \kappa W d^{3/2}K_Y\psi_{\mathbf{X}}\psi_\varepsilon^2 c^{-1}\left(\sqrt{\frac{d}{D}} + \frac{5}{3}\kappa W\right) \tag{5.23}$$

Proof. Let $\overline{w}_{k,m} := \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]$, and similarly $\overline{w}^A_{k,m} := \mathbb{E}\left[w_k(A\mathbf{X};A\mathbf{z}_m)\right]$ and $\overline{w}^\perp_{k,m} := \mathbb{E}\left[w_k(A^\perp\mathbf{X};A^\perp\mathbf{z}_m)\right]$ and recall from Lemma 12 that $\overline{w}_{k,m} \geq \frac{3}{7}$. We will also use the result from Lemma 14 that $\mathrm{var}(w_k(\mathbf{x};\mathbf{z}_m)) \leq k^2W^2D^2$ and similar bounds on $\mathrm{var}(w_k(A\mathbf{x};A\mathbf{z}_m))$ and $\mathrm{var}(w_k(A^\perp\mathbf{x};A^\perp\mathbf{z}_m))$ in Corollary 9, and the result of Lemma 3 that bounds the distance between $\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{r}_m\mathbf{X}\right]$ and $\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]$.

$$\begin{aligned}\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{r}_m\mathbf{X}\right]\| &\leq \left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]\right\| + C_R\tau^3\sqrt{\frac{D\log_2(2N))^3}{N}} \\ &\leq \left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]\right\| + C_RD^{-1/2}\tau^3(\log_2(2N))^{3/2}\end{aligned}$$

Turning our attention to the first term, letting $\overline{f}_{m,k} = \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]^{-1}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\right]$,

$$\begin{aligned}
\left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]\right\| &= \left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\mathbf{X}\right] - \overline{f}_{m,k}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\right]\right\| \\
&\leq \left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})A\mathbf{X}\right]\right\| + \left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})(A^{\perp}\mathbf{X})\right]\right\| + \left\|\overline{f}_{m,k}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\right]\right\| \\
&\leq \left\|\mathbb{E}\left[(w_k(A\mathbf{X};A\mathbf{z}_m) - \overline{w}^A_{k,m})f(A\mathbf{X})(A\mathbf{X})\right]\right\| + \left\|\overline{f}_{m,k}\mathbb{E}\left[(w_k(\mathbf{X};\mathbf{z}_m) - \overline{w}_{k,m})\mathbf{X}\right]\right\| \\
&\quad + \left\|\mathbb{E}\left[(w_k(A\mathbf{X};A\mathbf{z}_m) - \overline{w}^A_{k,m})f(A\mathbf{X})\right]\mathbb{E}\left[(w_k(A^{\perp}\mathbf{X};A^{\perp}\mathbf{z}_m) - \overline{w}^{\perp}_{k,m})A^{\perp}\mathbf{X}\right]\right\| \\
&\leq \sqrt{\mathbb{E}\left[(w_k(A\mathbf{X};A\mathbf{z}_m) - \overline{w}^A_{k,m})^2\right]\mathbb{E}\left[\|f(A\mathbf{X})(A\mathbf{X})\|^2\right]} \\
&\quad + \overline{f}_{m,k}\sqrt{\mathbb{E}\left[(w_k(\mathbf{X};\mathbf{z}_m) - \overline{w}_{k,m})^2\right]\mathbb{E}\left[\|\mathbf{X}\|^2\right]} \\
&\quad + \sqrt{\mathbb{E}\left[(w_k(A\mathbf{X};\mathbf{z}_m) - \overline{w}^A_{k,m})^2\right]\mathbb{E}\left[f(A\mathbf{X})^2\right]\mathbb{E}\left[(w_k(A^{\perp}\mathbf{X};\mathbf{z}_m) - \overline{w}^{\perp}_{k,m})^2\right]\mathbb{E}\left[\|A^{\perp}\mathbf{X}\|^2\right]} \\
&\leq kW\psi_{\mathbf{X}}\tau K_Y d^2\sqrt{2\log_2(2N)} + \overline{f}_{m,k}kWD^{3/2}\psi_{\mathbf{X}}\sqrt{2} + k^2W^2D^{3/2}\psi_{\mathbf{X}}\tau K_Y d^{3/2}\sqrt{2\log_2(2N)} \\
&\leq \kappa D^{-1/2}d^{3/2}W\psi_{\mathbf{X}}\tau K_Y\sqrt{2\log_2(2N)}\left(D^{-1/2}\sqrt{d} + \kappa W\right) + \kappa\overline{f}_{m,k}W\psi_{\mathbf{X}}\sqrt{2D}
\end{aligned}$$

Additionally, $\overline{f}_{m,k}$ itself is not too large, with

$$\begin{aligned}
\left|\overline{f}_{m,k}\right| = \left|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]^{-1}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\right]\right| &\leq \frac{7}{3}\mathbb{E}\left[(w_k(A\mathbf{X};\mathbf{z}_m) - \overline{w}^A_{k,m})f(A\mathbf{X})\right] \\
&\leq \frac{7}{3}kWd^{3/2}\tau K_Y\sqrt{\log_2(2N)} \leq \frac{2}{3}\kappa WD^{-1}d^{3/2}\tau K_Y\sqrt{\log_2(2N)}
\end{aligned}$$

yielding the desired bound. □

Next, we show that $\left\langle \hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m, \nu\right\rangle$ satisfies a sub-Gaussian concentration inequality.

Lemma 2. *Taking assumptions 1-4 and 6, and conditioning on the event (C.5) that bounds the sample second moment of* $\mathbf{X}$ *and the high probability bound on* r *in Corollary 8, the random variable* $\hat{S}\hat{\mathbf{p}}_m$ *is sub-Gaussian around* $\mathbf{u}_m$*:*

$$\mathbb{P}\left[\left\|\hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m\right\| > t\right] \leq 2\exp\left(-cN\frac{t^2}{75\tau^6 DK_R^2\psi_{\mathbf{X}}^2(\log_2 2N)^3}\right) \tag{5.24}$$

Proof. Indeed, first note that $w_k(\mathbf{X};\mathbf{z}_m)\tilde{r}_{m,k}\mathbf{X}$ is a sub-Gaussian vector, as $\tilde{r}_{m,k}$ is

bounded by $K_R + |\overline{r}_m|$ and $w_k(\mathbf{X}; \mathbf{z}_m) \leq 1$:

$$\begin{aligned}
\|w_k(\mathbf{X}; \mathbf{z}_m)\tilde{r}_{m,k}\mathbf{X}\|_\psi &\leq \sup_{\|\nu\|=1} \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{|w_k(\mathbf{X}; \mathbf{z}_m)\tilde{r}_{m,k}\langle \mathbf{X}, \nu\rangle|^2}{t^2}\right)\right] \leq 2 \right\} \\
&\leq \sup_{\|\nu\|=1} \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{\left(\tau^3 K_R (\log_2 2N)^{3/2} + |\overline{r}_m|\right)^2 |\langle \mathbf{X}, \nu\rangle|}{t}\right)\right] \leq 2 \right\} \\
&= \left(\tau^3 K_R (\log_2 2N)^{3/2} + |\overline{r}_m|\right) \psi_{\mathbf{X}}
\end{aligned}$$

Additionally, $\overline{r}_m$ is not too large:

$$|\overline{r}_m| = \left| \frac{\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m) r_i}{\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m)} \right| \leq 4\tau^3 K_R (\log_2 2N)^{3/2}$$

and hence

$$\begin{aligned}
&\leq \left(\tau^3 K_R (\log_2 2N)^{3/2} + 4\tau^3 K_R (\log_2 2N)^{3/2}\right) \psi_{\mathbf{X}} \\
&= 5\tau^3 K_R \psi_{\mathbf{X}} (\log_2 2N)^{3/2}
\end{aligned}$$

Thus, via Hoeffding's inequality, for any unit vector ν

$$\begin{aligned}
\mathbb{P}\left[\left|\left\langle \hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m, \nu \right\rangle\right| > t\right] &= \mathbb{P}\left[\left|\left\langle \frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i; \mathbf{z}_m)\tilde{r}_{m,i}\mathbf{x}_i - \mathbf{u}_m, \nu \right\rangle\right| > t\right] \\
&\leq 2\exp\left(-cN\frac{t^2}{25\tau^6 K_R^2 \psi_{\mathbf{X}}^2 (\log_2 2N)^3}\right)
\end{aligned}$$

Via Corollary 6, we have the desired bound on the norm. □

These lemmata allow us to prove Theorem 3:

Proof of Theorem 3. We will show this through the tail bound definition of the sub-Gaussian norm in Lemma 5, applied to the decomposition:

$$\hat{\mathbf{p}}_m - \mathbf{u}_m = \hat{\mathbf{p}}_m - \hat{S}^{-1}\mathbf{u}_m + \hat{S}^{-1}\mathbf{u}_m - \mathbf{u}_m = \hat{S}^{-1}(\hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m) + \hat{S}^{-1}(I - \hat{S})\mathbf{u}_m$$

Indeed, for arbitrary unit vector ν

$$\begin{aligned}
\mathbb{P}\left[\left\|\hat{S}^{-1}(\hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m)\right\| > t\right] &\leq \mathbb{P}\left[\left\|\hat{S}\hat{\mathbf{p}}_m - \mathbf{u}_m\right\| > \frac{t}{2}\right] \\
&\leq 2\exp\left(-cN\frac{t^2}{300\tau^6 D K_R^2 \psi_{\mathbf{X}}^2 (\log_2 2N)^3}\right)
\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left[\left\|(I-\hat{S})\hat{S}^{-1}\mathbf{u}_m\right\| > t\right] &\leq \mathbb{P}\left[\left\|I-\hat{S}\right\| > \frac{t}{\left\|\hat{S}^{-1}\mathbf{u}_m\right\|}\right] \\ &\leq \mathbb{P}\left[\left\|I-\hat{S}\right\| > \frac{t\sqrt{D}}{2V\tau^3(\log_2(2N))^{3/2}}\right] \\ &\leq 2\exp\left(-cN\frac{t^2}{8V^2\tau^6(\log_2(2N))^3C_S^2\psi_{\mathbf{X}}^4}\right)\end{aligned}$$

Combining these two bounds, we thus have

$$\begin{aligned}\mathbb{P}\left[\|\hat{\mathbf{p}}_m-\mathbf{u}_m\| > t\right] &\leq \mathbb{P}\left[\left|\left\langle \hat{S}^{-1}(\hat{S}\hat{\mathbf{p}}_m-\mathbf{u}_m),\nu\right\rangle\right| > \frac{t}{2}\right] + \mathbb{P}\left[\left|\left\langle (I-\hat{S})\hat{S}^{-1}\mathbf{u}_m,\nu\right\rangle\right| > \frac{t}{2}\right] \\ &\leq 2\exp\left(-cN\frac{t^2}{1200\tau^6DK_R^2\psi_{\mathbf{X}}^2(\log_2 2N)^3}\right) \\ &\quad + 2\exp\left(-cN\frac{t^2}{32V^2\tau^6(\log_2(2N))^3C_S^2\psi_{\mathbf{X}}^4}\right) \\ &\leq 4\exp\left(-cN\frac{t^2}{16\tau^6D\psi_{\mathbf{X}}^2(\log_2(2N))^3\left(75K_R^2+2V^2D^{-1}C_S^2\psi_{\mathbf{X}}^2\right)}\right)\end{aligned}$$

Given the trivial bound of $\mathbb{P}\left[\|\hat{\mathbf{p}}_m-\mathbf{u}_m\| > t\right] \leq 1$, we may take the square root of our bound to restore the coefficient, and so

$$\mathbb{P}\left[\|\hat{\mathbf{p}}_m-\mathbf{u}_m\| > t\right] \leq 2\exp\left(-cN\frac{t^2}{32\tau^6D\psi_{\mathbf{X}}^2(\log_2(2N))^3\left(75K_R^2+2V^2D^{-1}C_S^2\psi_{\mathbf{X}}^2\right)}\right)$$

as desired. □

The $N^{-1/2}$ coefficient of the sub-Gaussian norm is exactly what will eventually yield the $N^{-1/2}$ convergence rate, through the bound it gives us on the singular values of the error matrix.

Corollary 1. *Let T_2 be the $M \times D$ matrix with rows η_m as defined in Theorem 3. Then, with probability greater than $1-2\exp(-t^2)$,*

$$\|T\| \leq t\sqrt{\frac{\tau^6MD(\log_2 2N)^3\psi_\eta^2\log_2 2M}{N}}$$

Proof. From Theorem 3,

$$\mathbb{P}\left[\max_{m=1,\dots M}\|\eta_m\| > t\right] \leq 2\exp\left(-\frac{Nt^2}{\tau^6D(\log_2 2N)^3\psi_\eta^2\log_2 2M}\right)$$

Since by the triangle inequality $\|T\| \leq \sqrt{M}\,\|T\|_{2,\infty}$, we have the desired bound. $\square$

Note that because the $\hat{\mathbf{p}}_m$ are computed with the same $\mathbf{x}_i$, they are not independent and thus we cannot use Theorem 6 to bound the singular values here. One could modify MPLS to partition such that each $\hat{\mathbf{p}}_m$ is computed on independent partitions, however this would introduce an extra factor of $\sqrt{M}$ and overall complicate the analysis.

5.2.3 Approximation error of $\mathbf{u}_m$

Now that we have shown that the estimate $\hat{\mathbf{p}}_m$ concentrates around $\mathbf{u}_m$, we must bound the difference between $\mathbf{u}_m$ and $\mathbf{p}_m$. It is important to recognize that enforcing the condition $\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\right] = 0$ results in a non-linear dependence on $\mathbf{x}$, and so the connection between $\mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\tilde{r}_{m,i}\mathbf{x}_i\right]$ and $\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]$ is not immediately clear. Additionally, r_i is not actually a draw from the random variable $f(A\mathbf{X})+\varepsilon$, but instead is constructed from an estimate of β. Using the bound on the error of the estimate $\hat{\beta}$ in Corollary 7, we show that $\mathbf{u}_m$ is close to $\mathbf{p}_m$.

Lemma 3. *Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N/2}$, $\mathcal{S}' = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{N/2}$ be independent samples drawn from $\mathbf{X} \times Y$, which satisfy assumptions 1-4, and let k satisfy assumption 6, and suppose $N \geq \max(D^3, 32)$. Assume $\hat{\beta}$ minimizes the least squares problem (5.2) on $\mathcal{S}'$, and recall the definitions*

$$\begin{aligned} r_i &= y_i - \left\langle \hat{\beta}, \mathbf{x}_i \right\rangle \\ \tilde{r}_{m,i} &= r_i - \left(\frac{2}{N}\sum_{j=1}^{N/2} w_k(\mathbf{x}_j;\mathbf{z}_m)\right)^{-1} \frac{2}{N}\sum_{j=1}^{N/2} w_k(\mathbf{x}_j;\mathbf{z}_m) r_j \\ \tilde{f}_{m,k}(A\mathbf{X}) &= f(A\mathbf{X}) - \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]^{-1}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m) f(A\mathbf{X})\right] \end{aligned}$$

Then, with probability greater than $1 - 10\exp\left(-\tau^2\right)$, $\tau > \tau_\varepsilon(\log_2 2N)^{-1/2}$, the bounds

of Lemma 10 and Corollary 8 hold, and

$$\left\|\mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\tilde{r}_{m,i}\mathbf{x}_i\right] - \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right]\right\| \leq C_R\tau^3\sqrt{\frac{D(\log_2 2N)^3}{N}} \tag{5.25}$$

where

$$C_R := 3\tau_\varepsilon^{-2}\sqrt{\frac{C_\beta}{2}} + 64K_R\psi_{\mathbf{X}}\left(\sqrt{\frac{32\pi\kappa^2W^2}{7c}} + \frac{6}{\sqrt{N}}\right) + \frac{6}{7\sqrt{N}}\left(K_R\psi_{\mathbf{X}} + \tau_\varepsilon^{-2}K_Y\psi_{\mathbf{X}}\sqrt{d}\right) \tag{5.26}$$

and C_β is defined in (C.11).

Proof. We prove this in several parts, noting the decompositions

$$\begin{aligned}
&\mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\tilde{r}_{m,i}\mathbf{x}_i\right] \\
&\quad= \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)r_i\mathbf{x}_i\right] - \mathbb{E}\left[\frac{\left(\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)r_i\right)\left(\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\mathbf{x}_i\right)}{\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)}\right] \\
&\quad= A_1 - A_2
\end{aligned} \tag{5.27}$$

$$\begin{aligned}
&\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\mathbf{X}\right] \\
&\quad= \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\mathbf{X}\right] - \frac{\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\right]\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\right]}{\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]} \\
&\quad= B_1 - B_2
\end{aligned} \tag{5.28}$$

Let $A = A_1 - A_2$ and $B = B_1 - B_2$. We will show that $\|A_1 - B_1\|$ and $\|A_2 - B_2\|$ are small. First, we expand A_1, noting via the least squares normal equations,

$$\hat{\beta} = \left(\frac{2}{N}\sum_{i=1}^{N/2}(\mathbf{x}_i')(\mathbf{x}_i')^T\right)^{-1}\frac{2}{N}\sum_{i=1}^{N/2} y_i'\mathbf{x}_i' = \beta + \left(\frac{2}{N}\sum_{i=1}^{N/2}(\mathbf{x}_i')(\mathbf{x}_i')^T\right)^{-1}\frac{2}{N}\sum_{i=1}^{N/2}(f(A\mathbf{x}_i') + \varepsilon_i')\mathbf{x}_i'$$

Thus,

$$\begin{aligned}
r_i &= y_i - \left\langle\hat{\beta},\mathbf{x}_i\right\rangle \\
&= \left\langle\beta - \hat{\beta},\mathbf{x}_i\right\rangle + f(A\mathbf{x}_i) + \varepsilon_i - \left(\frac{2}{N}\sum_{i=1}^{N/2}(\mathbf{x}_i')(\mathbf{x}_i')^T\right)^{-1}\frac{2}{N}\sum_{i=1}^{N/2}(f(A\mathbf{x}_i') + \varepsilon_i')\left\langle\mathbf{x}_i',\mathbf{x}_i\right\rangle
\end{aligned} \tag{5.29}$$

We can decompose A_1 as follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m) r_i \mathbf{x}_i\right] &= \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\left(y_i - \left\langle\hat{\beta},\mathbf{x}_i\right\rangle\right)\mathbf{x}_i\right] \\
&= \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m) f(A\mathbf{x}_i)\mathbf{x}_i\right] \\
&\quad + \mathbb{E}\left[\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)\left(\left\langle\beta-\hat{\beta},\mathbf{x}_i\right\rangle + \varepsilon_i\right)\mathbf{x}_i\right] \\
&= B_1 + \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\mathbf{X}^T\right]\mathbb{E}\left[\beta-\hat{\beta}\right]
\end{aligned}$$

Now that $\mathbb{E}\left[\beta-\hat{\beta}\right]$ is small, as by Corollary 7,

$$\left\|\mathbb{E}\left[\beta-\hat{\beta}\right]\right\| \le \mathbb{E}\left[\left\|\beta-\hat{\beta}\right\|\right] \le \sqrt{\frac{2C_\beta\tau^2 D\log_2 2N}{N}}$$

And, by the sample well-behavedness condition (C.5), $\left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\mathbf{X}^T\right]\right\| \le \left\|\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]\right\| \le \frac{3}{2}$, and so

$$\|A_1 - B_1\| \le 3\tau\sqrt{\frac{C_\beta D\log_2 2N}{2N}} \tag{5.30}$$

Bounding $A_2 - B_2$ is more complicated, as the non-linear dependence among the $\mathbf{x}_j$ complicates the analysis. First, we note that

$$A_2 = \mathbb{E}\left[\frac{\frac{4}{N^2}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)^2 r_i\mathbf{x}_i + \frac{4}{N^2}\sum_{1\le i\ne j\le N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)w_k(\mathbf{x}_j;\mathbf{z}_m) r_i\mathbf{x}_j}{\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)}\right]$$

By independence of the $\mathbf{x}_i$'s, we may simplify this to

$$A_2 = \mathbb{E}\left[\frac{\frac{2}{N}w_k(\mathbf{x}_1;\mathbf{z}_m)^2 r_1\mathbf{x}_1 + \frac{N-2}{N}w_k(\mathbf{x}_1;\mathbf{z}_m)w_k(\mathbf{x}_2;\mathbf{z}_m) r_1\mathbf{x}_2}{\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m)}\right] \tag{5.31}$$

Define $\omega := \frac{2}{N}\sum_{j=3}^{N/2} w_k(\mathbf{x}_j,\mathbf{z}_m)$ and

$$g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2) := \frac{2}{N}w_k(\mathbf{x}_1;\mathbf{z}_m)^2 r_1\mathbf{x}_1 + \frac{N-2}{N}w_k(\mathbf{x}_1;\mathbf{z}_m)w_k(\mathbf{x}_2;\mathbf{z}_m) r_1\mathbf{x}_2$$

Thus, noting that $\omega \perp\!\!\!\perp g$

$$A_2 = \mathbb{E}\left[\frac{g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2)}{\frac{2}{N}\sum_{j=1}^{N/2} w_k(\mathbf{x}_j;\mathbf{z}_m)}\right] = \mathbb{E}\left[g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2)\mathbb{E}\left[\left(\omega+\frac{2t}{N}\right)^{-1}\middle|\sum_{i=1}^{2} w_k(\mathbf{x}_i;\mathbf{z}_m) = t\right]\right]$$

We now decompose A_2 as

$$\begin{aligned} A_2 &= \mathbb{E}\left[g(\mathbf{x}_1, r_1, \mathbf{x}_2, r_2)\mathbb{E}\left[\left(\omega + \frac{2t}{N}\right)^{-1} - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1} \middle| \sum_{i=1}^{2} w_k(\mathbf{x}_i; \mathbf{z}_m) = t\right]\right] \\ &\quad + \frac{\mathbb{E}\left[g(\mathbf{x}_1, r_1, \mathbf{x}_2, r_2)\right]}{\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]} \\ &= A_{2,1} + A_{2,2} \end{aligned} \tag{5.32}$$

$A_{2,1}$ is small, since $\left(\omega + \frac{2t}{N}\right)^{-1}$ cannot be too different from $\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1}$. Indeed, by Taylor's theorem, there is a random variable ξ with the same range as ω such that

$$\left(\omega + \frac{2t}{N}\right)^{-1} - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1} = -\frac{1}{\left(\xi + \frac{2t}{N}\right)^2}\left(\omega - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right] + \frac{2t}{N}\right)$$

Thus,

$$\mathbb{E}\left[\left(\omega + \frac{2t}{N}\right)^{-1} - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]^{-1}\right] \leq \left(\max_{\xi} \xi^{-2}\right) \mathbb{E}\left[\left|\omega - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right] + \frac{2t}{N}\right|\right]$$

From Lemma 13, the sample well-behavedness conditions force $\omega + \frac{2t}{N} \geq \frac{1}{4}$. Since $t \leq 2$, we have $\omega \geq \frac{1}{4} - \frac{4}{N}$. With the assumption $N \geq 32$, we have $\omega \geq \frac{1}{8}$, and thus $\xi^{-2} \leq 64$.

Additionally, we shall decompose

$$\omega - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right] + \frac{2t}{N} = (\omega - \mathbb{E}\left[\omega\right]) + \frac{2}{N}\left(2\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right] + t\right)$$

Now, by Lemma 14 (and the size of k in Assumption 6), $\operatorname{var}\left(w_k(\mathbf{X}; \mathbf{z}_m)\right) \leq k^2 W^2 D^2$. Since $w_k(\mathbf{X}; \mathbf{z}_m)$ is bounded between 0 and 1, we have from Corollary 5 that

$$\left\|w_k(\mathbf{X}; \mathbf{z}_m) - \mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z}_m)\right]\right\|_{\psi} \leq kWD\sqrt{2}.$$

Thus, by Hoeffding's inequality,

$$\mathbb{P}\left[\left|\omega - \mathbb{E}\left[\omega\right]\right| > t\right] \leq 2\exp\left(-\frac{c(N-4)t^2}{4k^2W^2D^2}\right)$$

and thus

$$\mathbb{E}\left[\left|\omega - \mathbb{E}\left[\omega\right]\right|\right] \leq \sqrt{\frac{4\pi k^2 W^2 D^2}{c(N-4)}} \leq \sqrt{\frac{4\pi \kappa^2 W^2 \psi_{\mathbf{X}}^2}{c(N-4)}}$$

Hence,

$$\mathbb{E}\left[\left(\omega+\frac{2t}{N}\right)^{-1}-\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]^{-1}\right]\leq 64\left(\sqrt{\frac{4\pi\kappa^2W^2}{c(N-4)}}+\frac{6}{N}\right) \tag{5.33}$$

To complete the bound on $A_{2,1}$, we bound $\mathbb{E}\left[\|\|g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2)\|\|\right]$ through the bounds on r in Corollary 8 and the moment bound on the sub-Gaussian variable $\|\mathbf{X}\|$ via Corollary 6:

$$\begin{aligned}\mathbb{E}\left[\|\|g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2)\|\|\right]&\leq\frac{2}{N}\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)^2\left\|r_1\mathbf{x}_1\right\|\right]\\&\quad+\frac{N-2}{N}\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)\left|r_1\right|\right]\mathbb{E}\left[w_k(\mathbf{x}_2;\mathbf{z}_m)\left\|\mathbf{x}_2\right\|\right]\\&\leq\tau^3K_R(\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D}\end{aligned} \tag{5.34}$$

From (5.33) and (5.34), we thus have

$$\|A_{2,1}\|\leq 64\tau^3K_R(\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D}\left(\sqrt{\frac{4\pi\kappa^2W^2}{c(N-4)}}+\frac{6}{N}\right) \tag{5.35}$$

Finally, $A_{2,2}$ is approximately equal to B_2. Indeed,

$$\begin{aligned}A_{2,2}-B_2&=\frac{\mathbb{E}\left[g(\mathbf{x}_1,r_1,\mathbf{x}_2,r_2)\right]}{\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right]}-B_2\\&=\frac{2}{N}\left(\frac{\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)^2r_1\mathbf{x}_1\right]}{\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)\right]}-B_2\right)\end{aligned}$$

We thus upper bound the two terms,

$$\begin{aligned}\left\|\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)^2r_1\mathbf{x}_1\right]\right\|&\leq\tau^3K_R(\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D}\\\left\|\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)f(A\mathbf{X})\right]\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\mathbf{X}\right]\right\|&\leq\tau K_Y\psi_{\mathbf{X}}\sqrt{dD\log_2 2N}\end{aligned}$$

and hence, recalling that $\mathbb{E}\left[w_k(\mathbf{x}_1;\mathbf{z}_m)\right]\geq\frac{3}{7}$,

$$\|A_{2,2}-B_2\|\leq\frac{6}{7N}\left(\tau^3K_R(\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D}+\tau K_Y\psi_{\mathbf{X}}\sqrt{dD\log_2 2N}\right) \tag{5.36}$$

We thus combine the bounds (5.30), (5.35), and (5.36) to yield

$$\begin{aligned}\|A-B\| &\leq \|A_1 - B_1\| + \|A_{2,1}\| + \|A_{2,2} - B_2\| \\ &\leq 3\tau\sqrt{\frac{C_\beta D \log_2 2N}{2N}} \\ &\quad + 64\tau^3 K_R (\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D}\left(\sqrt{\frac{4\pi\kappa^2 W^2}{c(N-4)}} + \frac{6}{N}\right) \\ &\quad + \frac{6}{7N}\left(\tau^3 K_R (\log_2 2N)^{3/2}\psi_{\mathbf{X}}\sqrt{D} + \tau K_Y \psi_{\mathbf{X}}\sqrt{dD\log_2 2N}\right)\end{aligned}$$

Collecting terms, using the assumption $\tau > \tau_\varepsilon (\log_2 2N)^{-1/2}$ to homogenize the bound in τ, and taking $N \geq 32$, we have

$$\begin{aligned}\|A-B\| \leq \tau^3\sqrt{\frac{D(\log_2 2N)^3}{N}}\Bigg[3\tau_\varepsilon^{-2}\sqrt{\frac{C_\beta}{2}} + 64K_R\psi_{\mathbf{X}}\left(\sqrt{\frac{32\pi\kappa^2W^2}{7c}} + \frac{6}{\sqrt{N}}\right) \\ + \frac{6}{7\sqrt{N}}\left(K_R\psi_{\mathbf{X}} + \tau_\varepsilon^{-2}K_Y\psi_{\mathbf{X}}\sqrt{d}\right)\Bigg]\end{aligned}$$

as desired. □

5.2.4 The Intrinsic Solution

We now wish to show that the asymptotic solutions $\mathbf{p}_m$ have full d rank and appreciable norm. Throughout this section, we take Assumptions 1-6, although only the low-dimensional aspects will be needed. The results in this section will also hold with probability $1 - 6\exp(-\tau^2)$ through the boundedness conditions of Lemma 10. Due to the separability of w and A-independence of the probability distribution, we have from (5.19):

$$\mathbf{p}_m = \mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]\mathbb{E}\left[w_k(A\mathbf{X}; A\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})A\mathbf{X}\right]$$

Moreover, Lemma 3 allows us to focus on this quantity, as it is close to the expectation of the empirical quantities we compute.

In this section all computations are done in the intrinsic space—no terms besides $\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]$ depend[6] on D.

[6]The assumptions on k in Assumption 6 also force $k \approx D^{-1}$ in the other sections, however here we only require the low-dimensional bounds (5.11) and (5.12)

We also recall from Assumption 2 and the separability of w_k that $\mathbb{E}[A\mathbf{X}] = 0$, that $\mathbb{E}\left[w_k(A\mathbf{X}; A\mathbf{z}_m)\tilde{f}_{m,k}(A\mathbf{X})\right] = 0$, and that $\mathbb{E}[f(A\mathbf{X})A\mathbf{X}] = 0$.

Finally, we will need the structural assumption 5 on Q from which we will derive the lower bound on the singular values of $\hat{P}$. We will be thus looking at minimizers of

$$\mathbb{E}\left[\left(f(A\mathbf{X})\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 - \langle A\mathbf{X}, \mathbf{q}(\mathbf{z}_m)\rangle\right)^2\right]$$

By the normal equations,

$$\mathbf{q}(\mathbf{z}_m) = \mathbb{E}\left[A\mathbf{X}(A\mathbf{X})^T\right]^{-1}\mathbb{E}\left[f(A\mathbf{X})\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 A\mathbf{X}\right].$$

To proceed, we introduce the following functions of r, with their reliance on $\mathbf{z}_m$ suppressed from their notation:

$$\begin{aligned}
\mathbf{c}(r) &:= \mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)f(A\mathbf{X})A\mathbf{X}] - \frac{\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)f(A\mathbf{X})]\,\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)A\mathbf{X}]}{\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)]} \\
\mathbf{v}(r) &:= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m)\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 f(A\mathbf{X})A\mathbf{X}\right] \\
\mathbf{b}(r) &:= \frac{\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)f(A\mathbf{X})]\,\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)A\mathbf{X}]}{\mathbb{E}[w_r(A\mathbf{X}; A\mathbf{z}_m)]}
\end{aligned}$$

Note that $\mathbf{c}(k) = \mathbb{E}\left[w(A^{\perp}\mathbf{X})\right]^{-1}\mathbf{p}_m$, $\mathbf{c}(0) = 0$, and $\mathbf{v}(0) = \mathbf{q}(\mathbf{z}_m)$. We shall now see that $\mathbf{c}(k) \approx k\mathbf{v}(0)$, by first noting the differential relationships among our parameters:

$$\frac{d}{dr}\mathbf{c}(r) = -\mathbf{b}'(r) - \mathbf{v}(r)$$

We will proceed by proving that $\mathbf{b}'(r)$ is small for all r in consideration and that $\mathbf{v}(r)$ is approximately constant. This will thus show that the derivative of $\mathbf{c}(r)$ is approximately constant and thus $\mathbf{c}(r)$ grows at an approximately linear rate.

First, that $\mathbf{b}'(r)$ is small.

Lemma 4. *For all $0 \le r \le k$, $\|\mathbf{b}'(r)\| \le 3r$.*

Proof. This is just a long series of bounds, since $\mathbf{b}'(r)$ has six types of terms in it:

$$\begin{aligned} B_1 &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m)\right]^{-1} & B_1' &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m)\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2\right] \\ B_2 &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m) f(A\mathbf{X})\right] & B_2' &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m)\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 f(A\mathbf{X})\right] \\ \mathbf{B}_3 &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m) A\mathbf{X}\right] & \mathbf{B}_3' &= \mathbb{E}\left[w_r(A\mathbf{X}; A\mathbf{z}_m)\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2 A\mathbf{X}\right], \end{aligned}$$

as

$$\mathbf{b}'(r) = rB_1^2 B_1' B_2 \mathbf{B}_3 - rB_1 B_2' \mathbf{B}_3 - rB_1 B_2 \mathbf{B}_3'. \tag{5.37}$$

Note that $\left\| \left\|A(\mathbf{X}-\mathbf{z}_m)\right\| \right\|_\psi \leq (\psi_A + K_A)\sqrt{d}$, hence

$$\begin{aligned} B_1 &\leq \mathbb{E}\left[1 - r\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^2\right]^{-1} \leq 2 \\ B_2 &= \mathbb{E}\left[(w_r(A\mathbf{X}; A\mathbf{z}_m) - \overline{W}_r) f(A\mathbf{X})\right] \leq rW_A d V_f \\ \|\mathbf{B}_3\| &= \left\|\mathbb{E}\left[(w_r(A\mathbf{X}; A\mathbf{z}_m) - \overline{W}_r) A\mathbf{X}\right]\right\| \leq rW_A d^{3/2} \psi_A \sqrt{2} \\ B_1' &\leq 2d\left(\psi_A + K_A\right)^2 \\ B_2' &\leq V_f \mathbb{E}\left[\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^4\right]^{1/2} \leq 4V_f d\left(\psi_A + K_A\right)^2 \\ \|\mathbf{B}_3'\| &\leq \mathbb{E}\left[\left\|A(\mathbf{X}-\mathbf{z}_m)\right\|^4\right]^{1/2} \mathbb{E}\left[\left\|A\mathbf{X}\right\|^2\right]^{1/2} \leq (2d)^{3/2}\left(\psi_A + K_A\right)^2 \psi_A \end{aligned}$$

Notice thus that each term of (5.37) has at least a factor of r^2 in it: one factor of r from the coefficient and another from either B_2 or $\mathbf{B_3}$. With the duo of bounds (5.12) on k in Assumption 6, noting that the bound on $B_1' B_2 \mathbf{B}_3$ has a factor of r^3 and the bounds on $B_1 B_2' \mathbf{B}_3$ and $B_1 B_2 \mathbf{B}_3'$ are equal, each term is thus bounded by r, and so

$$\|\mathbf{b}'(r)\| \leq 3r$$

□

With these bounds, we can prove the main result of this section:

Theorem 4. *For all* $0 \leq r \leq k$,

$$\|\mathbf{c}(r) + r\mathbf{v}(0)\| \leq \left(\frac{3 + C_1 d^{\frac{5}{2}}}{2}\right) r^2 \tag{5.38}$$

with C_1 *defined in* (5.39).

Proof. We prove this by appealing to the fundamental theorem of calculus, first noting that $\mathbf{v}(s)$ is approximately constant for small s:

$$\begin{aligned}
\|\mathbf{v}(r) - \mathbf{v}(0)\| &= \left\|\mathbb{E}\left[(1 - w_r(A\mathbf{X}; A\mathbf{z}_m)) \|A(\mathbf{X} - \mathbf{z}_m)\|^2 f(A\mathbf{X}) A\mathbf{X}\right]\right\| \\
&\leq rV_f \mathbb{E}\left[\|A(\mathbf{X} - \mathbf{z}_m)\|^{10}\right]^{4/10} \mathbb{E}\left[\|A\mathbf{X}\|^{10}\right]^{1/10} \\
&\leq rV_f (10d)^{5/2} (\psi_A + K_A)^4 \psi_A \\
&=: C_1 d^{\frac{5}{2}} r
\end{aligned} \tag{5.39}$$

Thus,

$$\begin{aligned}
\|\mathbf{c}(r) + r\mathbf{v}(0)\| &= \left\|\int_0^r c'(s) + \mathbf{v}(0) ds\right\| \\
&= \left\|\int_0^r -\mathbf{b}'(s) - \mathbf{v}(s) + \mathbf{v}(0) ds\right\| \\
&\leq \int_0^r \|\mathbf{b}'(s)\| \, ds + \int_0^r \|\mathbf{v}(0) - \mathbf{v}(s)\| \, ds \\
&\leq \int_0^r 3s ds + \int_0^r C_1 d^{\frac{5}{2}} s ds \\
&= \left(\frac{3 + C_1 d^{\frac{5}{2}}}{2}\right) r^2
\end{aligned}$$

□

Corollary 2. *The asymptotic solution* $\mathbf{p}_m$ *is approximately equal to* $\mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] k\mathbf{q}(\mathbf{z}_m)$, *as*

$$\left\|\mathbf{p}_m + k\mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] \mathbf{v}(0)\right\| \leq \mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] \left(\frac{3 + C_1 d^{\frac{5+p}{2}}}{2}\right) k^2 \tag{5.40}$$

This quadratic bound, since k is small, is very useful. Now, we can use Lemma 6 on the $M \times d$ matrix $k\mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] Q$ with rows $k\mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] \mathbf{q}(\mathbf{z}_m) = k\mathbb{E}\left[w_k(A^\perp \mathbf{X})\right] \mathbf{v}(0)$.

Corollary 3. *Recalling bound (5.11) of Assumption 6, the matrix P with rows $\mathbf{p}_m$ has large singular values:*

$$\left\|P^{-1}\right\|^{-1} \geq \frac{k\lambda_Q\sqrt{M}}{4}$$

Proof. Let $\gamma = \frac{1}{2}$. Since by Lemma 12, we have $\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right] \geq \frac{1}{\sqrt{6}}$ and by (5.11)

$$k \leq \frac{3\sigma_d(Q)^2}{8\left\|Q\right\|_F^2\left(3+C_1d^{\frac{5}{2}}\right)}\left\|\mathbf{q}(\mathbf{z}_m)\right\|$$

The bound in (5.40) yields

$$\begin{aligned}\left\|\mathbf{p}_m - k\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]\mathbf{q}(\mathbf{z}_m)\right\| &\leq \left(\frac{3+C_1d^{\frac{5}{2}}}{2}\right)\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]k^2 \\ &\leq \frac{3\sigma_d(Q)^2}{16\left\|Q\right\|_F^2}\left\|\mathbf{q}(\mathbf{z}_m)\right\|\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]k\end{aligned}$$

By Lemma 6 with $\gamma = \frac{1}{2}$

$$\left\|P^{-1}\right\|^{-2} \geq \frac{3}{8}\sigma_d(k\mathbb{E}\left[w_k(A^\perp\mathbf{X})\right]Q)^2\,,$$

and so, since $\mathbb{E}\left[w_k(A^\perp\mathbf{X};A^\perp\mathbf{z}_m)\right] \geq \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right] \geq 6^{-1/2}$ by Lemma 12,

$$\left\|P^{-1}\right\|^{-1} \geq \frac{1}{4}k\sigma_d(Q) \geq \frac{k\lambda_Q}{4}\sqrt{M}.$$

□

Appendix A

Influence of Ambient Dimension in Contour Regression

[LL20] gives a comprehensive proof of the $\sqrt{N}$-consistency of GCR, in which they show that the angle between the true dimension reduction space Φ and the GCR regressed one $\hat{\Phi}$ is bounded as follows

$$\mathbb{E}\left[\|\text{Proj}_{\hat{\Phi}} - \text{Proj}_{\Phi}\|^2\right] \leq \frac{C_7}{N}$$

which then allows for accurate regression with polynomial splines $\hat{f}$

$$\mathbb{E}\left[\left\|\hat{f}(\mathbf{X}) - f(\mathbf{X})\right\|^2_{L^2(\rho)}\right] \leq \frac{16C_8}{N} + 2C\left(\frac{\max(\sigma^2 + 2C_8N^{-1}, 2M^2 + 4C_8N^{-1})\log N}{N}\right)^{\frac{2p}{2p+d}}$$

However, we shall see that the constants C_7 and C_8 are potentially exponential in $\sqrt{D}$, and avoiding that requires $\frac{N}{\log N} \gtrsim D^{D/2}$.

A.1 Dependence of C_7 and C_8 on Assumption 1

The issue revolves around Assumption 1 of [LL20]:

Assumption 7. *There exist $\alpha_{thresh} > 0$, $C_0 > 0$ and a function $\phi : \mathbb{R} \to \mathbb{R}$ satisfying $\phi(\alpha) < C_0$ for $\alpha \in (0, \alpha_{thresh})$, such that: for any $\alpha \in (0, \alpha_{thresh})$ and unit vectors $\mathbf{v} \in \mathcal{S}_\Phi$ and $\mathbf{w} \in \mathcal{S}_\Phi^\perp$, the following hold:*

$$\text{var}\left[\mathbf{v}^T(\tilde{\mathbf{x}} - \mathbf{x})|V_f(\tilde{\mathbf{x}}, \mathbf{x}) \leq \alpha\right] \leq \phi(\alpha)$$

$$\text{var}\left[\mathbf{w}^T(\tilde{\mathbf{x}}-\mathbf{x})|V_f(\tilde{\mathbf{x}},\mathbf{x})\leq\alpha\right]\geq C_0$$

where

$$V_f(\tilde{\mathbf{x}},\mathbf{x})=\text{var}\left(f(\mathbf{z})|\mathbf{z}\in\ell(\tilde{\mathbf{x}},\mathbf{x})\right)$$

and $\ell(\tilde{\mathbf{x}},\mathbf{x})$ *is the line connecting* $\tilde{\mathbf{x}}$ *and* $\mathbf{x}$*.*

The authors note that Theorem 4.2 in [LZC05] shows that, for α sufficiently small, the assumption can hold (as the necessary inequality in variances is possible) for the model

$$y_i=g(\Phi^T\mathbf{x}_i)+\xi_i$$

These constants enter into the result through the value $C_0-\phi(\alpha+\alpha_0+3\sigma^2)$, which appears as a term in C_4 and C_6, which are propagated to C_7 and C_8. C_8, then, is the coefficient of the N^{-1} terms of the regression error in the main result. Let $\lambda=C_0-\phi(\alpha+\alpha_0+3\sigma^2)$, and we reproduce the values of the constants in terms of λ:

$$\begin{aligned}
C_4&=\lambda^2\\
C_5&=1152B^4\\
C_6&=64B^2\lambda\\
C_7&=\lambda^{-2}\left[C_5(\log(2D)+2D+1)+8\lambda^2+2^{15}B^4D\right]\\
C_8&=\lambda^{-2}dL_g^2B^2\left[C_5(\log(2D)+2D+1)+8\lambda^2+2^{15}B^4D\right]
\end{aligned}$$

As seen in the proof of Theorem 4.2, λ increases as α decreases to σ^2. However, it makes no claim as to the rate of this relationship. This is important, as α_0 and α are both potentially cursed terms, defined as

$$\alpha_0:=\max\left\{C_2\left(\frac{\log N}{N}\right)^{1/D},C_3\left(\frac{\log N}{N}\right)^{\frac{1}{D+2}}\right\}\tag{A.1}$$

$$\alpha:=4dL_g^2B^2\left(\frac{\log N}{N}\right)^{1/D}+a_0+3\sigma^2.\tag{A.2}$$

Thus if α and α_0 are required to be too small, this would enforce a curse of dimensionality on N. We shall see that in the single-index model, this is the case.

A.2 The Single-Index Model

Proposition 1. *In the single-index model* $f(\mathbf{x}) = c\,\langle \mathbf{x}, \nu \rangle$, $(\|\nu\| = 1)$, $y = f(\mathbf{x}) + \varepsilon$, $\mathbb{E}[\varepsilon^2] = \sigma^2$, $\mathbf{x}$ *uniformly distributed on the unit ball in* $\mathbb{R}^D$, *we must have either* $a_0 < D^{-1/2}$ *or* $C_7, C_8 \in \Omega\left(\exp\left(\sqrt{D}\right)\right)$.

Proof. First, we note what the requirement that $V_f(\mathbf{x}, \tilde{\mathbf{x}}) < \alpha$ entails. Let $\mathbf{m} = \frac{1}{2}(\mathbf{x}+\tilde{\mathbf{x}})$ and $t : -\frac{1}{2} \to \frac{1}{2}$, then

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{m} + t(\mathbf{x} - \tilde{\mathbf{x}}))\right] &= \mathbb{E}\left[f(\mathbf{m}) + tc\,\langle \nu, \mathbf{x} - \tilde{\mathbf{x}} \rangle\right] \\ &= f(\mathbf{m})\end{aligned}$$

Thus,

$$V_f(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{E}\left[(f - \mathbb{E}[f])^2\right] + \sigma^2 = \mathbb{E}\left[(f - f(\mathbf{m}))^2)\right] + \sigma^2$$

we compute

$$\begin{aligned}\mathbb{E}\left[(f(\mathbf{m} + t(\mathbf{x} - \tilde{\mathbf{x}})) - f(\mathbf{m}))^2\right] &= \mathbb{E}\left[c^{-2}t^2\,\langle \nu, \mathbf{x} - \tilde{\mathbf{x}} \rangle^2\right] \\ &= \frac{c^{-2}\,\langle \nu, \mathbf{x} - \tilde{\mathbf{x}} \rangle^2}{12}\end{aligned}$$

Thus, the requirement that

$$V_f(\mathbf{x}, \tilde{\mathbf{x}}) < \alpha$$

is equivalent to

$$\langle \nu, \mathbf{x} - \tilde{\mathbf{x}} \rangle^2 < \frac{12\,(\alpha - \sigma^2)}{c^2}$$

Now, we turn our attention to λ, which the difference between the orthogonal $(\mathbf{w} \perp \nu)$ and intrinsic $(\mathbf{v} = \nu)$ variances, thus

$$\lambda(\alpha) = \mathbb{E}\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 \,\middle|\, \langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^2\alpha\right] - \mathbb{E}\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 \,\middle|\, \langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^2\alpha\right]$$

By splitting on the condition $\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)$, we see that

$$\begin{aligned}
\lambda(\alpha) &= \Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 \geq 12c^{-2}(\alpha - \sigma^2) \,\middle|\, \langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right] \\
&\quad \cdot \mathbb{E}\left[\langle \mathbf{x}_1 - \tilde{\mathbf{x}}_1, \mathbf{w} \rangle^2 - \langle \mathbf{x}_2 - \tilde{\mathbf{x}}_2, \mathbf{v} \rangle^2 \,\middle|\, \begin{matrix} \langle \mathbf{x}_i - \tilde{\mathbf{x}}_i, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2), \\ \langle \mathbf{x}_i - \tilde{\mathbf{x}}_i, \mathbf{w} \rangle^2 \geq 12c^{-2}(\alpha - \sigma^2) \end{matrix}\right] \\
&\leq 8 \Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 \geq 12c^{-2}(\alpha - \sigma^2) \,\middle|\, \langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right] \\
&= 8 \frac{\Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 \geq 12c^{-2}(\alpha - \sigma^2)\right]}{\Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right]} \\
&\quad \cdot \Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2) \,\middle|\, \langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 \geq 12c^{-2}(\alpha - \sigma^2)\right] \\
&\leq 8 \left(\frac{1 - \Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right]}{\Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{v} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right]} \right)
\end{aligned}$$

Now, by spherical symmetry the probabilities in the numerator and denominator are equal. Additionally, by restricting to the case where both $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are in an equatorial strip,

$$\Pr\left[\langle \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{w} \rangle^2 < 12c^{-2}(\alpha - \sigma^2)\right] \geq \Pr\left[\langle \mathbf{x}, \mathbf{w} \rangle^2 < 3c^{-2}(\alpha - \sigma^2)\right]^2$$

Appealing to the concentration bounds for spherical caps ([BHK20], Theorem 2.7),

$$\begin{aligned}
\Pr\left[\langle \mathbf{x}, \mathbf{w} \rangle^2 < 3c^{-2}(\alpha - \sigma^2)\right]^2 &\geq \left(1 - \frac{2\exp\left(-\frac{3c^{-2}(\alpha-\sigma^2)(D-1)}{2}\right)}{\sqrt{(3c^{-2}(\alpha - \sigma^2))(D-1)}}\right)^2 \\
&\geq 1 - \frac{4\exp\left(-\frac{3c^{-2}(\alpha-\sigma^2)(D-1)}{2}\right)}{\sqrt{(3c^{-2}(\alpha - \sigma^2))(D-1)}}
\end{aligned}$$

Thus,

$$\begin{aligned}
\lambda(\alpha) &\leq \frac{32\exp\left(-\frac{3}{2}c^{-2}(\alpha - \sigma^2)(D-1)\right)}{\sqrt{(3c^{-2}(\alpha - \sigma^2)(D-1))} - 4\exp\left(-\frac{3}{2}c^{-2}(\alpha - \sigma^2)(D-1)\right)} \\
&\leq \frac{32\exp\left(-\frac{3}{2}c^{-2}(\alpha - \sigma^2)(D-1)\right)}{\sqrt{(3c^{-2}(\alpha - \sigma^2)(D-1))} - 4}
\end{aligned}$$

In order for this to not be exponentially small in $\sqrt{D}$, we require $c^{-2}(\alpha - \sigma^2) \leq \frac{1}{\sqrt{D}}$. However, by (A.2), this requires

$$c^{-2}\left(4dL_g^2B^2\left(\frac{\log N}{N}\right)^{1/D} + a_0 + 2\sigma^2\right) \leq \frac{1}{\sqrt{D}}$$

Even if σ^2 is small enough for this inequality to be possible, we must have $a_0 \leq D^{-1/2}$ to avoid the exponential dependence[1].

Returning to C_7 and C_8, this yields a factor of

$$\frac{\left(\sqrt{(3c^{-2}(\alpha-\sigma^2)(D-1))}-4\right)^2 \exp\left(3c^{-2}(\alpha-\sigma^2)(D-1)\right)}{1024}$$

Thus, if $a_0 > D^{-1/2}$, C_7 and C_8 have a factor of $\exp\left(\sqrt{D}\right)$.

□

Finally, the requirement that $a_0 \leq D^{-1/2}$ is a stringent one, as due to (A.1) it requires

$$\frac{N}{\log N} > (C_2 D)^{D/2}$$

It is worth noting further that, again due to concentration, $\lambda(\alpha) \lesssim D^{-1}$, and so even in the small α regime, C_7 and C_8 are cubic in D. Thus, we have as a corollary,

Corollary 4. *In the single-index model, one of the following must be true:*

- *C_7 and C_8 depend exponentially on $\sqrt{D}$*
- *$\frac{N}{\log N} \gtrsim D^{D/2}$*

The first option is more palatable, leaving the N^{-1} rate undisturbed. However, with a cursed constant, it will still require N very large for $\frac{C_7}{N}$ to begin being small.

[1] Of course, we actually need $a_0 \leq D^{-1}$, but this requirement allows for a more clear proof that something is wrong no matter what

Appendix B

Cube Regression

When the probability distribution is rotationally invariant, we must use supervised methods in order to determine the intrinsic subspace. However, if we can exploit existing structure, then it may be possible to determine the intrinsic subspace with higher accuracy and/or less restrictive assumptions. For the purposes of regression, determining the subspace with higher accuracy is somewhat academic, as the overall regression error will be dominated by the $N^{\frac{-2p}{2p+d}}$ minimax lower bound (Fact 7) that holds even for perfect subspace recovery. However, other data science purposes may wish to learn the subspace more accurately, and having a faster rate may help balance the various constants more favorably.

In dimension reduction literature, artificial experiments are often done on data that is uniformly distributed on a cube and where the intrinsic subspace is aligned with the faces of the cube. In this scenario, one could learn the orientation of the cube independent of the y data and then check each axis for whether it is independent of the function or not—for example, dividing the interval into slices and computing the variance of y in each slice, then taking the axes with smallest average sliced variance[1].

It is difficult, however, to determine the rotation of a cube in high dimensions. Principal component analysis fails in the uniform case, as the variance along any

[1]This is different from Sliced Average Variance Estimation, or SAVE, in that it is looking at the variance of y in slices of the $\mathbf{x}$, while SAVE looks at the variance of $\mathbf{x}$ in slices of the y.

projection is equal:

$$2^{-D} \int_{[-1,1]^D} \langle \mathbf{x}, \nu \rangle^2 \, dx = \sum_{i=1}^{D} \frac{1}{2} \int_{-1}^{1} \langle \nu, \mathbf{e}_i \rangle^2 \, x_i^2 dx_i = \frac{1}{3}$$

Moreover if $N < 2^D$, some vertices will not have points in them, which also makes naïve approaches difficult.

We introduce an algorithm that experimentally does fairly well at finding the orientation of a cube that exploits the fact that points in the vertices have large Euclidean norm, and Euclidean norm is preserved by rotations. Thus, if we select the points with largest norm and find a rotation that moves them close to vectors of ± 1, we will have recovered the rotation.

Algorithm 2. *Pick a parameter $K < N$ of points to regress, for example $K = N/D$. Sort the data $\mathbf{x}_i$ so that $\|\mathbf{x}_1\| \geq \|\mathbf{x}_2\| \geq \cdots \geq \|\mathbf{x}_K\|$ are the K datapoints with largest Euclidean norm, and let X^0 be the $K \times D$ matrix with $\mathbf{x}_i$ as rows. Then, for $t = 1, \ldots, K$*

1. *Let R_t be the $D \times D$ orthogonal matrix minimizing*

$$\left\| X_{:t}^{t-1} R_t - \operatorname{sgn} X_{:t}^{t-1} \right\|_F$$

   where $X_{:t}^{t-1}$ consists of the first t rows of X^{t-1},

2. *Then, let $X^t = X^{t-1} R_t$, so that the points are moved closer to their predicted vertices.*

The estimated rotation is then the product $R_1 R_2 \cdots R_K$.

We do not provide any theoretical analysis of this algorithm, nor is it intended to be optimal. Instead, we offer this as a proof of concept and as a method to provide semi-supervised experimental results. Consider two of the experiments done in Chapter 4: Radial Cosine (4.3) and L1 (4.4). For these examples, we lower the ambient D to 10, as this Cube Regression algorithm appears to have very poor results for large D[2].

[2]On the other hand, most experiments in dimension reduction papers have $D \leq 10$.

B.1 Cube Regression for L1 and Radial Cosine

For L1 (Experiment 4.4), let $\mathbf{X} \in [-1, 1]^{10}$ be distributed uniformly with $d = 4$, while for Radial Cosine (Experiment 4.3) $\mathbf{X} \in [-2, 2]^{10}$ is distributed uniformly with $d = 2$. In both cases, Gaussian noise with variance 5% of the empirical variance is added, we let $N = 200, 400, 800, 1600, 3200$, and in MPLS we discard the linear estimate and use only the slope perturbation estimates.

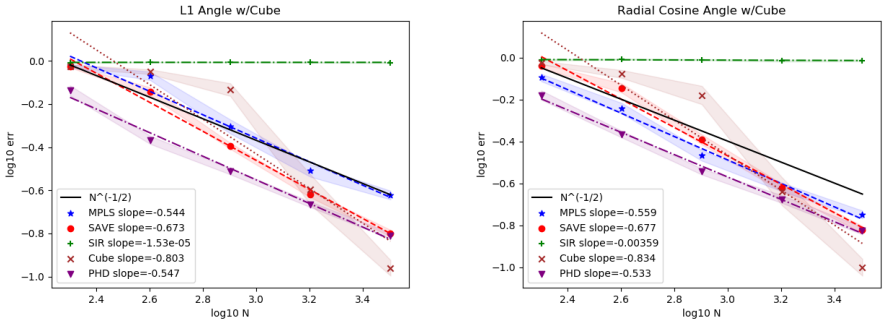

Figure B-1. Angle of regressed subspace, left L1, right Radial Cosine

The ability to use unsupervised learning methods frees us from the non-parametric minimax rate, and indeed we appear to achieve near N^{-1} rate. Interestingly, this is also faster than the convergence rate of random variables to their mean, which is $N^{-1/2}$. This is likely because there is no noise in the $\mathbf{X}$-values, and so there is a single correct answer that the algorithm attempts to find; recall from Chapter 1 that the noiseless minimax non-parametric lower bound converges much faster than the bound with noise. It is also possible that, due to the relatively high computational cost of this algorithm and the small value of D, this algorithm merely gets "lucky" as described in Section 2.7.

Appendix C

Proofs of Lemmas

C.1 Lemmas Independent of the Assumptions

We shall use the following properties of sub-Gaussian variables [Ver19]

Lemma 5 (Properties of Sub-Gaussian Random Variables)**.** *Let Z be a mean-zero sub-Gaussian random variable; the sub-Gaussian norm is thus*

$$\|Z\|_\psi := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{Z^2}{t^2}\right)\right] \leq 2\right\}$$

There exists an absolute constant c such that

1. *The tails of Z are bounded*

$$\mathbb{P}\left[|Z| > t\right] \leq 2\exp\left(-\frac{t^2}{\|Z\|_\psi^2}\right)$$

2. *The moments of Z are bounded for $p \geq 2$*

$$\mathbb{E}\left[|Z|^p\right]^{1/p} \leq \|Z\|_\psi \sqrt{p}$$

3. *The generalized Hoeffding's inequality holds for independent copies of Z:*

$$\mathbb{P}\left[\left|\sum_{i=1}^n Z_i - \mathbb{E}\left[Z\right]\right| \geq t\right] \leq 2\exp\left(\frac{-ct^2}{n\|Z\|_\psi^2}\right)$$

4. *Bernstein's inequality holds for independent copies of Z:*

$$\mathbb{P}\left[\left|\sum_{i=1}^n \left(Z_i^2 - \mathbb{E}\left[Z^2\right]\right)\right| \geq t\right] \leq 2\exp\left(-c\min\left(\frac{t^2}{n\|Z\|_\psi^4}, \frac{t}{\|Z\|_\psi^2}\right)\right)$$

It is worth noting that one could take distinct absolute constants in Hoeffding's inequality, and in Bernstein's inequality to achieve tighter bounds. We opt not to worry about tracking three different absolute constants, and instead take c to be large enough that the three inequalities hold.

Corollary 5. *Let Z be a random variable such that $|Z| \leq 1$ a.s. and $\mathbb{E}[Z^2] = \sigma^2$. Then $\|Z\|_\psi \leq \sigma\sqrt{2}$.*

Proof. Since Z is bounded by 1, for any $p \geq 2$,

$$\mathbb{E}[|Z|^p]^{1/p} \leq \mathbb{E}\left[Z^2\right]^{1/p} \leq \sigma$$

and, by Jensen's inequality, $\mathbb{E}[|Z|] \leq \sigma$. Thus,

$$\mathbb{E}\left[\exp\left(\frac{Z^2}{2\sigma^2}\right)\right] = \sum_{p=0}^{\infty} \frac{\mathbb{E}\left[\left(\frac{Z^2}{2\sigma^2}\right)^p\right]}{p!} \leq \sum_{p=0}^{\infty} \frac{1}{p!2^{2p}} \leq \frac{4}{3} \leq 2\,.$$

□

Corollary 6. *Let $\mathbf{Z}$ be a sub-Gaussian random vector in $\mathbb{R}^D$. Then $\|\mathbf{Z}\|$ is sub-Gaussian with $\|\|\mathbf{Z}\|\|_\psi \leq \sqrt{D}\,\|\mathbf{Z}\|_\psi$. In particular, if for any unit vector ν*

$$\mathbb{P}[|\langle \mathbf{Z}, \nu\rangle| > t] \leq 2\exp\left(-\frac{t^2}{K^2}\right)$$

then

$$\mathbb{P}[\|\mathbf{Z}\| > t] \leq 2\exp\left(-\frac{t^2}{3DK^2}\right).$$

Proof. From the tail bound, we know $\|\mathbf{Z}\|_\psi \leq \sqrt{3}K$, as

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\langle \mathbf{Z}, \nu\rangle^2}{3K^2}\right)\right] &= \int_0^\infty \mathbb{P}\left[\frac{\langle \mathbf{Z}, \nu\rangle^2}{3K^2} > s\right] ds \\
&\leq \int_0^\infty \min\left(1, 2\exp\left(-3\log s\right)\right) ds \\
&= 2^{1/3} + \int_{2^{1/3}}^\infty 2s^{-3} ds = 2^{1/3} + 2^{-2/3} < 2
\end{aligned}$$

Now, consider the Orlicz 1-norm, and note that $\|X\|_\psi = \sqrt{\|X^2\|_{\psi_1}}$:

$$\|X\|_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\left|\frac{X}{t}\right|\right)\right] \leq 2\right\}.$$

Since the Orlicz 1-norm satisfies the triangle inequality,

$$\| \|\mathbf{Z}\| \|_\psi = \sqrt{\left\| \|\mathbf{Z}\|^2 \right\|_{\psi_1}} \leq \sqrt{D} \|\mathbf{Z}\|_\psi$$

By Markov's inequality,

$$\mathbb{P}\left[\|\mathbf{Z}\| > t\right] \leq 2 \exp\left(-\frac{t^2}{3DK^2}\right)$$

as desired. □

In the proof of Theorem 1, we need to lower bound the singular values of the matrix of asymptotic solutions, P. This is done by showing that each row $\mathbf{p}_m$ is similar to another vector $\mathbf{q}_m$, which through Assumption 5 forms the rows of a matrix with large singular values. We use Wedin's theorem ([Wed72]) to bound the angle between $\hat{A}$ and A:

Theorem 5 (Wedin)**.** *Let $\hat{P} = P + T$, and let $P \approx U_1 \Sigma_1 V_1^T$ be the rank-d singular value approximation of P. Then, for any unitary-invariant norm,*

$$\left\|\sin\Theta(\hat{P}_d, P_d)\right\| \leq \frac{\max\left(\|TV_1\|, \left\|T^T U_1\right\|\right)}{\sigma_d(\hat{P}) - \sigma_{d+1}(P)}$$

where $(\hat{P})_d$ and P_d are the spaces spanned by the top d right singular vectors of $\hat{P}$ and P respectively. In particular, if P is rank d, then

$$\left\|\sin\Theta(\hat{P}_d, P_d)\right\| \leq \frac{\|T\|}{\sigma_d(\hat{P})}$$

Wedin's theorem requires us to lower bound the singular values of P. Here, we show that if the rows of two matrices are pairwise relatively close, then their smallest singular values must also be close. Indeed,

Lemma 6. *Let G be an $M \times d$ matrix with rows $\mathbf{g}_m$ and singular values $\sigma_1 \geq \cdots \geq \sigma_d$. Let $\tilde{\mathbf{g}}_m$ be a set of vectors such that $\|\tilde{\mathbf{g}}_m - \mathbf{g}_m\| \leq \gamma \left(\frac{3\sigma_d^2}{8\|G\|_F^2}\right) \|\mathbf{g}_m\|$, with $0 \leq \gamma < 1$. Then for any $\nu \in \mathbb{S}_d$,*

$$\sum_{m=1}^{M} \langle \tilde{\mathbf{g}}_m, \nu \rangle^2 \geq \frac{3}{4}(1-\gamma)\sigma_d^2$$

Proof. Let

$$\mathcal{M} := \left\{ m : \langle \mathbf{g}_m, \nu \rangle^2 \geq \frac{\sigma_d^2}{4 \|G\|_F^2} \|\mathbf{g}_m\|^2 \right\}$$

Then, by Lemma 7,

$$\begin{aligned}
\sum_{m=1}^{M} \langle \tilde{\mathbf{g}}_m, \nu \rangle^2 &\geq \sum_{m \in \mathcal{M}} \langle \tilde{\mathbf{g}}_m, v \rangle^2 = \sum_{m \in \mathcal{M}} \left(\langle \tilde{\mathbf{g}}_m - \mathbf{g}_m, \nu \rangle + \langle \mathbf{g}_m, v \rangle \right)^2 \\
&= \sum_{m \in \mathcal{M}} \left(\langle \tilde{\mathbf{g}}_m - \mathbf{g}_m, \nu \rangle^2 + 2 \langle \tilde{\mathbf{g}}_m - \mathbf{g}_m, \nu \rangle \langle \mathbf{g}_m, \nu \rangle + \langle \mathbf{g}_m, v \rangle^2 \right) \\
&\geq \sum_{m \in \mathcal{M}} \left(2 \langle \tilde{\mathbf{g}}_m - \mathbf{g}_m, \nu \rangle \langle \mathbf{g}_m, \nu \rangle + \langle \mathbf{g}_m, v \rangle^2 \right) \\
&\geq \sum_{m \in \mathcal{M}} \langle \mathbf{g}_m, \nu \rangle^2 - \sum_{m \in \mathcal{M}} 2 \|\tilde{\mathbf{g}}_m - \mathbf{g}_m\| \|\mathbf{g}_m\| \\
&\geq \frac{3}{4}\sigma_d^2 - 2\gamma \left(\frac{3\sigma_d^2}{8 \|G\|_F^2} \right) \sum_{m \in \mathcal{M}} \|\mathbf{g}_m\|^2 \geq \frac{3}{4}\sigma_d^2 - \gamma \left(\frac{3\sigma_d^2}{4} \right) = \frac{3}{4}(1-\gamma)\sigma_d^2
\end{aligned}$$

□

Lemma 7. *Let G be an $M \times d$ matrix with rows $\mathbf{g}_m$ and singular values $\sigma_1 \geq \cdots \geq \sigma_d$. Let $\nu \in \mathbb{S}_d$ be a unit vector, and $\mathcal{M} := \left\{ m : \langle \mathbf{g}_m, \nu \rangle^2 \geq \frac{\sigma_d^2}{4\|G\|_F^2} \|\mathbf{g}_m\|^2 \right\}$. Then*

$$\sum_{m \in \mathcal{M}} \langle \mathbf{g}_m, \nu \rangle^2 \geq \frac{3}{4}\sigma_d^2$$

Proof. The point here is that if G is full rank, then it cannot be the case that too many of the rows are orthogonal to any given vector. We use the fact that the smallest singular value is the minimum of the quadratic form $G^T G$ on the unit circle to deduce the result.

$$\begin{aligned}
\sum_{m \in \mathcal{M}} \langle \mathbf{g}_m, \nu \rangle^2 &= \sum_{m=1}^{M} \langle \mathbf{g}_m, \nu \rangle^2 - \sum_{m \notin \mathcal{M}} \langle \mathbf{g}_m, v \rangle^2 \geq \sigma_d^2 - \sum_{m \notin \mathcal{M}} \frac{\sigma_d^2}{4 \|G\|_F^2} \|\mathbf{g}_m\|^2 \\
&\geq \sigma_d^2 - \frac{\sigma_d^2}{4 \|G\|_F^2} \sum_{m=1}^{M} \|\mathbf{g}_m\|^2 = \sigma_d^2 - \frac{1}{4}\sigma_d^2 = \frac{3}{4}\sigma_d^2
\end{aligned}$$

□

C.2 Lemmas Relying on Assumptions 1-5

Lemma 8. *Under the assumptions on the distribution of $\mathbf{X}$ in Assumptions 2 and 3, the sample covariance matrix of a given partition $\hat{S}_j = \frac{2}{N} \sum_{i=1}^{N/2} \mathbf{x}_i \mathbf{x}_i^T$ concentrates*

around I:

$$\mathbb{P}\left[\left\|I-\hat{S}\right\| > \min\left(t, \frac{1}{2}\right)\right] \leq 2\exp\left(-c\frac{Nt^2}{2C_S^2\psi^4 D}\right) \tag{C.1}$$

where

$$C_S = \psi^{-2}\sqrt{\frac{c}{c_v D}} + C_v\sqrt{\frac{c}{\log 2}} \tag{C.2}$$

and c_v, C_v are the absolute constants in Theorem 6.

To prove this lemma, we will use the a theorem on the concentration of singular values of sub-Gaussian variables ([Ver12], Theorem 5.39/Remark 5.40).

Theorem 6. *Let T be a $K \times D$ matrix whose rows are independent sub-Gaussian vectors in $\mathbb{R}^D$ such that $\Sigma^{-1/2}T_i$ are isotropic, and let Ψ_T be the maximum of the sub-Gaussian norms of $\Sigma^{-1/2}T_i$. Then, for every $t > 0$, the following inequality holds with probability at least $1 - 2\exp\left(-c_v t^2\right)$:*

$$\left\|\frac{1}{K}T^T T - \Sigma\right\| \leq \|\Sigma\| \max\left(\delta, \delta^2\right)$$

where

$$\delta = \frac{C_v \Psi_T^2 \sqrt{D} + t}{\sqrt{K}}$$

and c_v and C_v are absolute constants.

Proof of Lemma 8. By Theorem 6, recalling from Assumption 2 that $\mathbf{X}$ is isotropic,

$$\mathbb{P}\left[\left\|I-\hat{S}\right\| > \max(\delta, \delta^2)\right] \leq 2\exp\left(-c_v t^2\right) \tag{C.3}$$

with $\delta = \frac{C_v\psi^2\sqrt{D}+t}{\sqrt{\frac{N}{2}}}$. Because we are only interested in bounds tighter than $\frac{1}{2}$, we may assume $\delta > \delta^2$. Thus, rearranging (C.3),

$$\mathbb{P}\left[\left\|I-\hat{S}\right\| > t\right] \leq 2\exp\left(-c_v\left(t\sqrt{\frac{N}{2}} - C_v\psi^2\sqrt{D}\right)^2\right) \tag{C.4}$$

Furthermore, since probabilities are bounded by 1 and, for all $s > 0$,

$$\max\left(s, (at-b)^2\right) \geq \max\left(s, \frac{s(at)^2}{(\sqrt{s}+b)^2}\right)$$

we can take $s = c_v^{-1}\log 2$ to yield the desired bound. □

Lemma 9 (Sample well-behavedness)**.** *Since* $\mathbf{X}$ *is sub-Gaussian, the following bounds are satisfied for both* $\mathcal{S}$ *and* $\mathcal{S}'$ *with high probability for arbitrary* ν *with* $\|\nu\| = 1$*:*

- *The bound on the second moments*

$$\left\| \frac{2}{N} \sum_{i=1}^{N/2} \mathbf{x}_i \mathbf{x}_i^T - I \right\| \leq \frac{1}{2} \tag{C.5}$$

  holds with probability greater than

$$1 - 2\exp\left(-c\frac{N}{8C_S^2\psi_{\mathbf{X}}^4 D}\right).$$

- *The bound on the mean*

$$\left\| \frac{2}{N} \sum_{i=1}^{N/2} \mathbf{x}_i \right\| \leq \frac{\sqrt{D}}{16} \tag{C.6}$$

  holds with probability greater than

$$1 - 2\exp\left(-c\frac{N}{1536\psi_{\mathbf{X}}}\right)$$

Define Ω_1 *to be the event of these two conditions holding across all partitions, then*

$$\mathbb{P}\left[\Omega\right] \leq 1 - 2\exp\left(-\frac{cN}{16\psi_{\mathbf{X}}^2 \max\left(C_S^2\psi_{\mathbf{X}}^2 D, 192\right)}\right)$$

Proof. The probability bound on the second moment (C.5) comes directly from Lemma 8.

The sample mean bound (C.6) arises from Hoeffding's inequality, noting that $\mathbb{E}\left[\mathbf{X}\right] = 0$, we have

$$\mathbb{P}\left[\left|\left\langle \frac{2}{N}\sum_{i=1}^{N/2} \mathbf{x}_i, \nu \right\rangle\right| > t\right] \leq 2\exp\left(-c\frac{Nt^2}{2\psi_{\mathbf{X}}^2}\right)$$

Thus, via Corollary 6,

$$\mathbb{P}\left[\left|\left\langle \frac{2}{N}\sum_{i=1}^{N/2} \mathbf{x}_i, \nu \right\rangle\right| > t\right] \leq 2\exp\left(-c\frac{Nt^2}{6D\psi_{\mathbf{X}}^2}\right)$$

Taking $t = \frac{1}{16}\sqrt{D}$ gives the desired bound. □

These events are not independent, however due to each being exponentially likely in N we may condition on the large probability event of them all occurring. Moreover, (C.5) combined with the invertibility of $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = I$ implies that the sample covariance matrix is invertible, with smallest eigenvalue greater than $1/2$.

We will also further condition on all $A\mathbf{x}_i$, $\mathbf{x}_i$, and y_i being bounded.

Lemma 10. *Let*

$$\tau_\varepsilon = \sqrt{\frac{c}{\psi_\varepsilon^2}}. \tag{C.7}$$

With probability greater than $1 - 6\exp\left(-\tau^2\right)$ *for* $\tau > \tau_\varepsilon(\log_2 2N)^{-1/2}$, *for all* $i = 1, \ldots, N$

$$\begin{aligned}
\|A\mathbf{x}_i\| &\leq \tau\psi_A\sqrt{\frac{d\log_2 2N}{c}} \\
&=: \tau K_{A\mathbf{X}}\sqrt{d\log_2 2N} && \text{(C.8)} \\
|y_i| &\leq 2C_f\left(\tau\psi_\varepsilon\sqrt{\frac{d\log_2 2N}{c}}\right) + \tau\max\left(\psi_{\mathbf{X}}, \psi_\varepsilon\right)\sqrt{\frac{d\log_2 2N}{c}}\left(2\|\beta\| + 1\right) \\
&=: \tau K_Y\sqrt{d\log_2 2N} && \text{(C.9)} \\
\|\mathbf{x}_i\| &\leq \tau\psi_{\mathbf{X}}\sqrt{\frac{D\log_2 2N}{c}} \\
&=: \tau K_{\mathbf{X}}\sqrt{D\log_2 2N} && \text{(C.10)}
\end{aligned}$$

Additionally, $|f(A\mathbf{x}_i)| \leq \tau K_Y\sqrt{d\log_2 2N}$.

Proof. By the sub-Gaussian probability bounds, since $\|A\mathbf{X}\|$ is sub-Gaussian with

norm bounded by $\psi_A\sqrt{d}$ and $\|\mathbf{X}\|$ is sub-Gaussian with norm bounded by $\psi_{\mathbf{X}}\sqrt{D}$:

$$\begin{aligned}
\mathbb{P}\left[\max_i \|A\mathbf{x}_i\| > s_1\right] &\leq 2N\exp\left(-c\frac{s_1^2}{d\psi_A^2}\right) \\
&\leq 2\exp\left(-c\frac{s_1^2}{d\psi_A^2\log_2 2N}\right) \\
\mathbb{P}\left[\max_i |y_i| > 2C_f s_1^{1\wedge p} + 2\|\beta\| s_2 + s_2 \middle| \max_i \|A\mathbf{x}_i\| < s_1\right] &\leq 4N\exp\left(-cs^2\min\left(4\psi_{\mathbf{X}}^{-2}, \psi_\varepsilon^{-2}\right)\right) \\
&\leq 2\exp\left(-c\frac{s^2\min\left(4\psi_{\mathbf{X}}^{-2}, \psi_\varepsilon^{-2}\right)}{2\log_2 2N}\right) \\
\mathbb{P}\left[\max_i \|\mathbf{x}_i\| > s_3\right] &\leq 2N\exp\left(-c\frac{s_3^2}{D\psi_{\mathbf{X}}^2}\right) \\
&\leq 2\exp\left(-c\frac{s_3^2}{D\psi_{\mathbf{X}}^2\log_2 2N}\right)
\end{aligned}$$

Thus, with probability greater than

$$1-2\exp\left(-c\frac{s_1^2}{d\psi_A^2\log_2 2N}\right)-2\exp\left(-c\frac{s^2\min\left(4\psi_{\mathbf{X}}^{-2}, \psi_\varepsilon^{-2}\right)}{2\log_2 2N}\right)-2\exp\left(-c\frac{s_3^2}{D\psi_{\mathbf{X}}^2\log_2 2N}\right)$$

we have $\max_i \|A\mathbf{x}_i\| \leq s_1$, $\max_i |y_i| \leq 2C_f s_1^{1\wedge p} + 2\|\beta\| s_2 + s_2$, and $\max_i \|\mathbf{x}_i\| \leq s_3$. We can thus lower bound this probability by

$$1 - 6\exp\left(-\frac{c}{\log_2 2N}\min\left(\frac{s_1^2}{d\psi_A^2}, \frac{2s^2}{\psi_{\mathbf{X}}^2}, \frac{s^2}{2\psi_\varepsilon^2}, \frac{s_3^2}{D\psi_{\mathbf{X}}^2}\right)\right)$$

The desired probability $1 - 6\exp(-\tau^2)$ is thus achieved by

$$\tau^2 = \frac{c}{\log_2 2N}\min\left(\frac{s_1^2}{d\psi_A^2}, \frac{2s^2}{\psi_{\mathbf{X}}^2}, \frac{s^2}{2\psi_\varepsilon^2}, \frac{s_3^2}{D\psi_{\mathbf{X}}^2}\right),$$

i.e. by choosing

$$s_1 = \tau\psi_A\sqrt{\frac{d\log_2 2N}{c}} \quad , \; s_2 = \tau\max(\psi_{\mathbf{X}}, \psi_\varepsilon)\sqrt{\frac{\log_2 2N}{c}} \quad , \; s_3 = \tau\psi_{\mathbf{X}}\sqrt{\frac{D\log_2 2N}{c}} \quad ,$$

and noting that $x^{1\wedge p} \leq x$ when $x \geq 1$. This yields the desired bounds. □

With these bounds and the sample well-behavedness condition, we can show sub-Gaussian concentration results for various quantities of interest.

Lemma 11. *Under the assumptions on the distribution of* $\mathbf{X}$ *and* Y *in Assumptions 1, 2, and 3 and the sample well-behavedness conditions in Lemma 9, with probability* $1 - 6\exp(-\tau^2)$ *Lemma 10 holds and each* $\hat{\beta}_m$ *concentrates around* β*; i.e.*

$$\mathbb{P}\left[\left\|\hat{\beta} - \beta\right\| > t\right] \leq 2\exp\left(-N\frac{t^2}{C_\beta \tau^2 D \log_2 2N}\right)$$

with

$$C_\beta := 32c^{-1}\psi_{\mathbf{X}}^2\left(3\tau^2 K_Y^2 d + 2\left\|\beta\right\|^2 C_S^2 \psi_{\mathbf{X}}^2 (\log_2 2N)^{-1}\right) \tag{C.11}$$

and C_S *defined in* (C.2)*.*

The consistency of linear least squares is well-known (for example, [GKKW02] Theorem 11.3), and the algorithm has been well studied in many configurations. We prove our own result here, however, because few results are available for our context of having random $\mathbf{x}_i$ and a response Y for which the "noise" $Y - \langle \beta, \mathbf{X}\rangle = f(A\mathbf{X}) + \varepsilon$ is poorly behaved (i.e. $\mathbb{E}\left[f(A\mathbf{X}) + \varepsilon \,|\mathbf{X}\right] \neq 0$). Proposition 3 of [Mou19] does prove a bound similar to Lemma 11, however the minor differences in problem definition make the result cumbersome to incorporate here.

Proof. By Lemma 8, we have sub-Gaussian concentration bounds on $\left\|\hat{S}\right\|$ yielding the bound (C.1) on $\left\|I - \hat{S}\right\|$. Next, bound the sub-Gaussian norm of $Y\mathbf{X}$,

$$\left\|Y\mathbf{X}\right\|_\psi \leq \tau K_Y \sqrt{d \log_2 2N}\psi_{\mathbf{X}}$$

Thus, via Hoeffding's inequality, for any unit vector ν,

$$\mathbb{P}\left[\left|\left\langle \frac{2}{N}\sum_{i=1}^{N/2} y_i'\mathbf{x}_i' - \mathbb{E}\left[Y\mathbf{X}\right], \nu\right\rangle\right| > t\right] \leq 2\exp\left(-cN\frac{t^2}{2\tau^2 K_Y^2 \psi_{\mathbf{X}}^2 d \log_2 2N}\right)$$

Noting that $\hat{S}\hat{\beta} = \frac{2}{N}\sum_{i=1}^{N/2} y_i'\mathbf{x}_i'$, we thus via Corollary 6 have the concentration

$$\mathbb{P}\left[\left\|\hat{S}\hat{\beta} - \beta\right\| > t\right] \leq 2\exp\left(-cN\frac{t^2}{6\tau^2 D K_Y^2 \psi_{\mathbf{X}}^2 d \log_2 2N}\right)$$

Then, we consider the decomposition

$$\hat{\beta} - \beta = \hat{S}^{-1}(\hat{S}\hat{\beta} - \beta) + (\hat{S}^{-1}\beta - \beta)$$

Both of these terms concentrate as follows:

$$\begin{aligned}
\mathbb{P}\left[\left\|\hat{S}^{-1}(\hat{S}\hat{\beta} - \beta)\right\| > s\right] &\leq \mathbb{P}\left[\left\|\hat{S}\hat{\beta} - \beta\right\| > \frac{s}{\left\|\hat{S}^{-1}\right\|}\right] \\
&\leq \mathbb{P}\left[\left\|\hat{S}\hat{\beta} - \beta\right\| > \frac{s}{2}\right] \\
&\leq 2\exp\left(-cN\frac{t^2}{12\tau^2 D K_Y^2 \psi_{\mathbf{X}}^2 d\log_2 2N}\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}\left[\left\|\hat{S}^{-1}\beta - \beta\right\| > s\right] &\leq \mathbb{P}\left[\left\|\hat{S}^{-1} - I\right\| \|\beta\| > s\right] \\
&\leq \mathbb{P}\left[\left\|I - \hat{S}\right\| > \frac{s}{\|\beta\| \left\|\hat{S}^{-1}\right\|}\right] \\
&\leq \mathbb{P}\left[\left\|I - \hat{S}\right\| > \frac{s}{2\|\beta\|}\right] \\
&\leq 2\exp\left(-cN\frac{t^2}{8\|\beta\|^2 C_S^2 \psi_{\mathbf{X}}^4 D}\right)
\end{aligned}$$

Combined, we have the desired result:

$$\mathbb{P}\left[\left\|\hat{\beta} - \beta\right\| > s\right] \leq 2\exp\left(-cN\frac{t^2}{32D\psi_{\mathbf{X}}^2\left(3\tau^2 K_Y^2 d\log_2 2N + 2\|\beta\|^2 C_S^2 \psi_{\mathbf{X}}^2\right)}\right)$$

□

Corollary 7.

$$\mathbb{E}\left[\left\|\hat{\beta} - \beta\right\|^2\right] \leq \frac{2C_\beta \tau^2 D\log_2 2N}{N}$$

where C_β is defined in (C.11)

With the concentration of $\hat{\beta}$ determined, we can expand our boundedness lemma to include bounds on $\left\|\hat{\beta} - \beta\right\|$ and $|r_i|$.

Corollary 8. *Conditioned on the sample well-behavedness events of Lemma 9, let* $\tau > \tau_\varepsilon(\log_2 2N)^{-1/2}$*, and assume* $N \geq D^2$*, then with probability* $1 - 8\exp(-\tau^2)$*:*

$$\begin{aligned}
\left\|\beta - \hat{\beta}\right\| &\leq \tau^2 \sqrt{\frac{16D\psi_{\mathbf{X}}^2 \log_2(2N)\left(3K_Y^2 cd + 2\left\|\beta\right\|^2 C_S^2 \psi_\varepsilon^2 \psi_{\mathbf{X}}^2\right)}{c^2 N}} \\
&=: \tau^2 K_{\hat{\beta}} \sqrt{\frac{\log_2(2N)}{D}} \qquad \text{(C.12)} \\
|r_i| &\leq \tau^3 \sqrt{\frac{\log_2 2N}{c}} \left(K_{\hat{\beta}} K_{\mathbf{X}} \sqrt{\log(2N)} + \frac{\psi_\varepsilon^2 \log 2N}{c}\left(2C_f \psi_A \sqrt{d} + \psi_\varepsilon\right)\right) \\
&=: \tau^3 K_R (\log_2 2N)^{3/2} \qquad \text{(C.13)}
\end{aligned}$$

Proof. With probability $1 - 6\exp(-\tau^2)$ we assume $|y_i|$ is bounded by $\tau K_Y \sqrt{d \log_2 2N}$ and $\|\mathbf{x}_i\| \leq \tau K_{\mathbf{X}} \sqrt{D}$. From Lemma 11, we have

$$\mathbb{P}\left[\left\|\hat{\beta} - \beta\right\| > s\right] \leq 2\exp\left(-cN_M \frac{s^2}{16D\psi_{\mathbf{X}}^2\left(3\tau^2 K_Y^2 d \log_2 2N + 2\left\|\beta\right\|^2 C_S^2 \psi_{\mathbf{X}}^2\right)}\right)$$

In order to get an exponent of $-\tau^2$, we take

$$s > \tau \sqrt{\frac{16D\psi_{\mathbf{X}}^2\left(3\tau^2 K_Y^2 d \log_2 2N + 2\left\|\beta\right\|^2 C_S^2 \psi_{\mathbf{X}}^2\right)}{cN_M}},$$

in particular, since $\tau > \psi_\varepsilon^{-1}\sqrt{\frac{c}{\log_2 2N}}$, we may choose

$$s = \tau^2 \sqrt{\frac{16D\psi_{\mathbf{X}}^2 \log_2(2N)\left(3K_Y^2 cd + 2\left\|\beta\right\|^2 C_S^2 \psi_\varepsilon^2 \psi_{\mathbf{X}}^2\right)}{c^2 N_M}}$$

yielding the desired bound after recalling that $N > D^2$.

Turning to the r_i,

$$\begin{aligned}
|r_i| &= \left|y_i - \left\langle \hat{\beta}, \mathbf{x}_i \right\rangle\right| \\
&= \left|f(A\mathbf{x}_i) + \varepsilon_i + \left\langle \beta - \hat{\beta}, \mathbf{x}_i \right\rangle\right| \\
&\leq \tau \sqrt{\frac{\log_2 2N}{c}}\left(2C_f \psi_A \sqrt{d} + \psi_\varepsilon + \left\|\beta - \hat{\beta}\right\| K_{\mathbf{X}} \sqrt{D}\right) \\
&\leq \tau \sqrt{\frac{\log_2 2N}{c}}\left(2C_f \psi_A \sqrt{d} + \psi_\varepsilon + \tau^2 K_{\mathbf{X}} \sqrt{D} K_{\hat{\beta}} \sqrt{\frac{\log_2(2N)}{D}}\right) \\
&\leq \tau^3 \sqrt{\frac{\log_2 2N}{c}}\left(K_{\hat{\beta}} K_{\mathbf{X}} \sqrt{\log(2N)} + \frac{\psi_\varepsilon^2 \log 2N}{c}\left(2C_f \psi_A \sqrt{d} + \psi_\varepsilon\right)\right)
\end{aligned}$$

as desired. □

C.3 Lemmas Supporting Assumption 6

The lemmas in this section will assume Assumptions 1-4, as well as statement (5.10) of Assumption 6.

Lemma 12. *Assume $\|\mathbf{z}\| \leq K_{\mathbf{z}}\sqrt{D}$ as in Assumption 5, and that k satisfies statement (5.10) of Assumption 6. Then for each m,*

$$\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right] \geq \frac{1}{\sqrt{6}}$$

Proof. We use the bound $e^{-x} \geq 1 - x$:

$$\begin{aligned}\mathbb{E}\left[w_k(\mathbf{X};\mathbf{z}_m)\right] &\geq 1 - k\mathbb{E}\left[\|\mathbf{X} - \mathbf{z}_m\|^2\right] = 1 - k\left(\mathbb{E}\left[\|\mathbf{X}\|^2\right] + \|\mathbf{z}_m\|^2\right) \\ &= 1 - kD - k\|\mathbf{z}_m\|^2 \geq \frac{3}{7} \geq \frac{1}{\sqrt{6}},\end{aligned}$$

where we use (5.10) in the second-to-last step[1]. □

We also will want a similar bound on the mean of the empirical weights.

Lemma 13. *Assume k satisfies statement (5.10) and condition on the sample well-behavedness event of Lemma 9. Then for each m,*

$$\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m) \geq \frac{1}{4}$$

Proof. Similar to before,

$$\begin{aligned}\frac{2}{N}\sum_{i=1}^{N/2} w_k(\mathbf{x}_i;\mathbf{z}_m) &\geq 1 - k\left(\frac{2}{N}\sum_{i=1}^{N/2}\|\mathbf{x}_i\|^2 + \|\mathbf{z}_m\|^2 - 2\langle\mathbf{x}_i,\mathbf{z}_m\rangle\right) \\ &\geq 1 - k\left(\frac{3}{2}\mathbb{E}\left[\|\mathbf{X}\|^2\right] + \|\mathbf{z}_m\|^2 + \frac{1}{8}\|\mathbf{z}_m\|\sqrt{D}\right) \geq \frac{1}{4}\end{aligned}$$

□

Lemma 14. *Assume k satisfies statement (5.10). Then, there is a constant W depending only on the sub-Gaussian norm of $\mathbf{X}$ such that*

$$\mathbb{E}\left[\left(w_k(\mathbf{X};\mathbf{z}) - \mathbb{E}\left[w_k(\mathbf{X};\mathbf{z})\right]\right)^2\right] \leq k^2W^2D^2$$

[1] $6^{-1/2}$ will be a more convenient number later, but is a little obnoxious to use in our bounds here. $\frac{3}{7} \approx 0.429$ while $6^{-1/2} \approx 0.408$.

To show this, we will need a technical lemma:

Lemma 15. *If a random variable B takes values on an interval I and $h : I \to \mathbb{R}$ is differentiable with $|h'(x)| \geq 1$ everywhere on I, then*

$$\operatorname{var}(B) \leq \operatorname{var}(h(B))$$

Proof. Indeed, by the mean value theorem for integrals there exists a number $b_1 \in I$ such that

$$h(b_1) = \mathbb{E}\left[h(B)\right]$$

Further, by the mean value theorem for derivatives, for any value of B there exists a number b_B such that

$$h(B) - h(b_1) = h'(b_B)(B - b_B)$$

Thus,

$$\begin{aligned}
\operatorname{var}(h(B)) &= \mathbb{E}\left[\left(h(B) - \mathbb{E}\left[h(B)\right]\right)^2\right] = \mathbb{E}\left[h'(b_B)^2\left(B - b_1\right)^2\right] \\
&\geq \mathbb{E}\left[\left(B - b_1\right)^2\right] = \operatorname{var}(B) + \left(b_1 - \mathbb{E}\left[B\right]\right)^2 \\
&\geq \operatorname{var}(B)
\end{aligned}$$

□

We now give the proof of Lemma 14:

Proof. To see that $\mathbb{E}\left[w_k(\mathbf{X}; \mathbf{z})\right]$ has small variance, we first note that

$$\operatorname{var}(w_k(\mathbf{X}; \mathbf{z})) \leq \operatorname{var}(k \left\|\mathbf{X} - \mathbf{z}\right\|^2)$$

which is true because $\frac{d}{dx}\log(x) \geq 1$ for all $x \in (0, 1]$ and thus follows from Lemma 15. Further, $\operatorname{var}(k \left\|\mathbf{X} - \mathbf{z}\right\|^2) = k^2 \operatorname{var}(\left\|\mathbf{X} - \mathbf{z}\right\|^2)$.

Noting that $\|\mathbf{X}\|$ is sub-Gaussian with sub-Gaussian norm bounded by $\psi_{\mathbf{X}}\sqrt{D}$,

$$\begin{aligned}
\text{var}\left(\|\mathbf{X}-\mathbf{z}\|^2\right) &\leq \mathbb{E}\left[\|\mathbf{X}-\mathbf{z}\|^4\right] \\
&\leq \mathbb{E}\left[(\|\mathbf{X}\|+\|\mathbf{z}\|)^4\right] \\
&= \mathbb{E}\left[\|\mathbf{X}\|^4\right] + 4\mathbb{E}\left[\|\mathbf{X}\|^3\right]\|\mathbf{z}\| + 6\mathbb{E}\left[\|\mathbf{X}\|^2\right]\|\mathbf{z}\|^2 + 4\mathbb{E}\left[\|\mathbf{X}\|\right]\|\mathbf{z}\|^3 + \|\mathbf{z}\|^4 \\
&\leq D^2\left(16\psi_{\mathbf{X}}^4 + 4\sqrt{27}K_{\mathbf{z}}\psi_{\mathbf{X}}^3 + 12K_{\mathbf{z}}^2\psi_{\mathbf{X}}^2 + 4K_{\mathbf{z}}^3\psi_{\mathbf{X}} + K_{\mathbf{z}}^4\right)
\end{aligned}$$

Let $W^2 = \left(16\psi_{\mathbf{X}}^4 + 4\sqrt{27}K_{\mathbf{z}}\psi_{\mathbf{X}}^3 + 12K_{\mathbf{z}}^2\psi_{\mathbf{X}}^2 + 4K_{\mathbf{z}}^3\psi_{\mathbf{X}} + K_{\mathbf{z}}^4\right)$ to yield the desired bound. □

Notably, since we expect $k \approx D^{-1}$, the weighted average of a random variable should not be too different from the unweighted average. This similarly holds for the variance of $w_k(A\mathbf{X}; A\mathbf{z})$ and $w_k(A^\perp\mathbf{X}; A^\perp\mathbf{z})$:

Corollary 9. *Under the conditions of Lemma 14, there exist constants W_A depending only on the sub-Gaussian norm of $A\mathbf{X}$, and $W_\perp$ depending only on the sub-Gaussian norm of $A^\perp\mathbf{X}$ such that*

$$\begin{aligned}
\mathbb{E}\left[\left(w_k(A\mathbf{X}; A\mathbf{z}) - \mathbb{E}\left[w_k(A\mathbf{X}; A\mathbf{z})\right]\right)^2\right] &\leq k^2 W_A^2 d^2 \\
\mathbb{E}\left[\left(w_k(A^\perp\mathbf{X}; A^\perp\mathbf{z}) - \mathbb{E}\left[w_k(A^\perp\mathbf{X}; A^\perp\mathbf{z})\right]\right)^2\right] &\leq k^2 W_\perp^2 (D-d)^2
\end{aligned}$$

It is worth noting that via Bernstein's inequality, assuming independence of the components of $\mathbf{X}$, the bound in Lemma 14 can be improved to k^2W^2D, potentially changing the value of W.

Bibliography

[BDK+17] Benedikt Bauer, Luc Devroye, Michael Kohler, Adam Krzyżak, and Harro Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

[BGM08] Gregory Beylkin, Jochen Garcke, and Martin J. Mohlenkamp. Multivariate regression and machine learning with sums of separable functions, 2008.

[BHK20] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

[Coo00] R. Dennis Cook. Save: a method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29(9-10):2109–2121, 2000.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

[DJS08] Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(53):1647–1678, 2008.

[ER09] M.J. Evans and J.S. Rosenthal. *Probability and Statistics: The Science of Uncertainty*. Macmillan Learning, 2009.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer, 2002.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HR78] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[JLT09] Anatoli B. Juditsky, Oleg V. Lepski, and Alexandre B. Tsybakov. Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3):1360–1404, 06 2009.

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[LL20] Hao Liu and Wenjing Liao. Learning functions varying along an active subspace, 2020.

[LMV20] Alessandro Lanteri, Mauro Maggioni, and Stefano Vigogna. Conditional regression for single-index models. 02 2020.

[Loy18] Joshua Loyal. sliced. https://joshloyal.github.io/sliced/, 2018.

[LPW+17] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In

Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6232–6240, Red Hook, NY, USA, 2017. Curran Associates Inc.

[LZ07] Yingxing Li and Li-Xing Zhu. Asymptotics for sliced average variance estimation. *The Annals of Statistics*, 35(1):41 – 69, 2007.

[LZC05] Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression: A general approach to dimension reduction. *Ann. Statist.*, 33(4):1580–1616, 08 2005.

[Mou19] Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv: Statistics Theory*, 2019.

[PD09] Adragni Kofi P. and Cook R. Dennis. Sufficient dimension reduction and prediction in regression. *Phil. Trans. R. Soc. A.*, 2009.

[Rud76] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

[SMP96] Lee Sigelman, Jarol B. Manheim, and Susannah Pierce. Inside dopes?: Pundits as political forecasters. *Harvard International Journal of Press/Politics*, 1(1):33–50, 1996.

[SS05] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis. Princeton Univ. Press, Princeton, NJ, 2005.

[Ste81] Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135 – 1151, 1981.

[Ver12] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

[Ver19] Roman Vershynin. High-dimensional probability. 2019.

[Wed72] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[WJ03] Qi-Hua Wang and Bing-Yi Jing. Empirical likelihood for partial linear models. *Annals of the Institute of Statistical Mathematics*, 55:585 – 595, 2003.

[WM97] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[XTLZ02] Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.